Technological University Dublin

ARROW@TU Dublin

Research Papers

51st Annual Conference of the European Society for Engineering Education (SEFI)

2023-10-10

The Stability Of Pre-Enrolment Prediction Of Academic Achievement: Criterion-Referencing Versus Norm-Referencing

Jolan HANSSENS
KU Leuven, Belgium;Leuven Engineering and Science Education Center (LESEC);Engineering Technology Education Research (ETHER), jolan.hanssens@kuleuven.be

Carolien VAN SOOM
KU Leuven, Belgium;Leuven Engineering and Science Education Center (LESEC), carolien.vansoom@kuleuven.be

Greet LANGIE
KU Leuven, Belgium;Leuven Engineering and Science Education Center (LESEC);Engineering Technology Education Research (ETHER), greet.langie@kuleuven.be

Follow this and additional works at: https://arrow.tudublin.ie/sefi2023_respap

Part of the Engineering Education Commons

Recommended Citation

Hanssens, J., Van Soom, C., & Langie, G. (2023). The Stability Of Pre-Enrolment Prediction Of Academic Achievement: Criterion-Referencing Versus Norm-Referencing. European Society for Engineering Education (SEFI). DOI: 10.21427/NH2E-WS39

This Conference Paper is brought to you for free and open access by the 51st Annual Conference of the European Society for Engineering Education (SEFI) at ARROW@TU Dublin. It has been accepted for inclusion in Research Papers by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, vera.kilshaw@tudublin.ie.

This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 4.0 International License.

The stability of pre-enrolment prediction of academic achievement: criterion-referencing versus norm-referencing

J. Hanssens [1]
KU Leuven, Leuven Engineering and Science Education Center (LESEC), Engineering Technology Education Research (ETHER), Faculty of Science, Faculty of Engineering Technology
B-3000 Leuven, Belgium
http://orcid.org/0000-0002-9282-9451

C. Van Soom
KU Leuven, Leuven Engineering and Science Education Center (LESEC), Faculty of Science
B-3000 Leuven, Belgium
https://orcid.org/0000-0001-7677-0931

G. Langie
KU Leuven, Leuven Engineering and Science Education Center (LESEC), Engineering Technology Education Research (ETHER), Faculty of Engineering Technology
B-3000 Leuven, Belgium
http://orcid.org/0000-0002-9061-6727

Conference Key Areas*: Recruitment and Retention of Engineering students*
Keywords*: Low-stakes assessment, norm-referencing, criterion-referencing, predictive validity, classification*

[1] *Corresponding Author*

J. Hanssens

jolan.hanssens@kuleuven.be

ABSTRACT

Positioning tests are organized in Flanders for prospective STEM students. They provide a low-stakes opportunity to assess their level of starting competences before enrolment. Predictive validity for subsequent academic achievement is an important quality measure of these positioning tests. However, the content of the tests varies over the years. This could be problematic for making accurate predictions based on data from previous years. Therefore, the objective of this study is to compare the stability over time of the predictions of academic achievement using either criterion-referenced (absolute grading) or norm-referenced (relative grading) positioning test grades of engineering and science students.

Comparisons of classifications over six academic years yielded various results (n=1258). For the engineering students, all predictions where unstable in those academic years when the tests were held online due to Covid-19 measures, and when positioning test participation became obligatory. However, in the years when aforementioned special events were absent, norm-referencing yielded the most stable prediction. For the science students, norm-referencing yielded a stable prediction over all six academic years, and criterion-referencing yielded a stable prediction when the tests were not held online. This clearly suggests that the implementation of norm-referencing in positioning tests may lead to more accurate predictions of academic achievement over time, regardless of changes in test content, despite the current use of criterion-referencing in practice.

1 INTRODUCTION

Positioning tests are a low-stakes opportunity to assess starting competencies before the start of higher education for prospective students in a program in Science, Engineering, Technology or Mathematics (STEM). The tests are organized in the summer holidays between the end of Secondary Education and the start of Higher Education. This allows prospective students to remedy any shortcomings in starting competencies before their first semester starts or even reconsider their study choice, in case of a low score (Vandewalle, and Callens 2013, 1-2). Note that these low scorers are not prohibited from entering the study program, as is the case with a high-stakes entrance exam. STEM programs in Flanders have open admission to anyone with a secondary degree and there is no centralized exam at the end of secondary education. Positioning tests are an attempt at solving the resulting issue of heterogeneity of academic preparedness of Flemish freshmen STEM students. Research on predictive validity for academic achievement of positioning tests generally compares different parts of the tests, or different predictors (Pinxten et al. 2019, 45-66; Vanderoost et al. 2014, 1-8; 2015, 1-8; Van den Broeck 2019, 989-1007). Such research focuses on *which* predictors exist, but not *how* to use them in actual predictions. This study aims to address that gap, and to practically improve the positioning test procedure.

One prominent issue with predictions of academic achievement based in positioning tests is the stability of the prediction over multiple academic years. There is always a need to categorize pseudo-continuous data of positioning test scores in order to determine cut-off scores for providing feedback to students. In general terms, this comes downs to the question of what grade does a student need to pass the test? This question can be answered based on historical data, i.e. in order to have such a chance to obtain such academic achievement, a student needs at least such a score, based on data from previous academic years. However, an issue with such statements is that considerable variation between academic years could arise, either between (i) the level and content of the problems on the test, (ii) or between the level of competencies of the cohorts of students taking the test. Yet, the accuracy of such statements is essential for providing adequate feedback to students. Therefore, this study investigates the stability of classifications of academic achievement based on positioning test scores over six academic years, 2016-2017 to 2021-2022. The focus is on the programme of Engineering Technology (ET), as well as the cluster of programmes Chemistry, Biology, Biochemistry and biotechnology, Geography and Geology (CBBGG).

The ET positioning test contained 20 mathematics problems and 10 text problems from 2016-2017 until 2019-2020. From 2020-2021 onwards, the text problems were omitted and the mathematics part was expanded and split into 10 basic mathematics problems and 15 standard level mathematics problems. The former have the specific goal of identifying students with a high risk of low academic achievement and are of a lower difficulty than the latter, which are similar to the mathematics problems of 2016-2017 – 2019-2020. The CBBGG positioning test contained 20 mathematics problems, 10 text problems and 10 chemistry problems from 2016-2017 until 2019-2020. In this test as well, the text problems were omitted and the mathematics part was expanded to 10 basic and 15 standard level mathematic problems from 2020-2021 onwards. The text problems of the ET test and of the CBBGG test were different, but they remained the same over the years. The mathematics problems of the ET test were the same as the CBBGG test in each year, but they varied each year. Finally, the chemistry problems of the CBBGG test varied over the years as well. Additionally, participation to the positioning test became obligatory in 2021-2022 (meaning that students had to take the test in order to enrol, but they did not need a passing grade), and both the ET and CBBGG tests were held online in 2020-2021 due to Covid-19 restrictions. Both these ‘special events’ are potential threats to the stability of the classification, as they potentially changed the composition of the participating cohort (voluntary versus obligatory participation) and test taking behaviour.

The aim of this study is to compare criterion-referencing and norm-referencing in terms of the stability of their prediction of academic achievement. Criterion-referencing, criterion-referenced grading, or absolute grading is comparing the students’ skill against a predetermined standard, often half of the maximal score. Norm-referencing, norm-referenced grading or relative grading, on the other hand, means comparing the skill of the student to that of their peers. Criterion-referencing based predictions can be hypothesized as more robust against changes in participant population and test taking behaviour, while norm-referencing based predictions can be hypothesized as more robust against changes in test composition. Note that positioning test composition has changed on two levels: changes *of* test parts (i.e. entire parts were added and omitted) and changes *within* test parts (i.e. the problems within some parts changed each year). Currently, positioning tests use criterion-referencing for determining cut-offs.

The research question of this study is: does either criterion- or norm-referencing of positioning test (partial) scores yield a more stable classification of academic achievement over the academic years 2016-2017 until 2021-2022 for ET and CBBGG students?

2 METHODOLOGY

In total, 1258 students participated in the positioning test for ET or CBBGG in the six academic years between 2016-2017 and 2021-2022 and subsequently enrolled in the corresponding study program at KU Leuven, (see table 1).

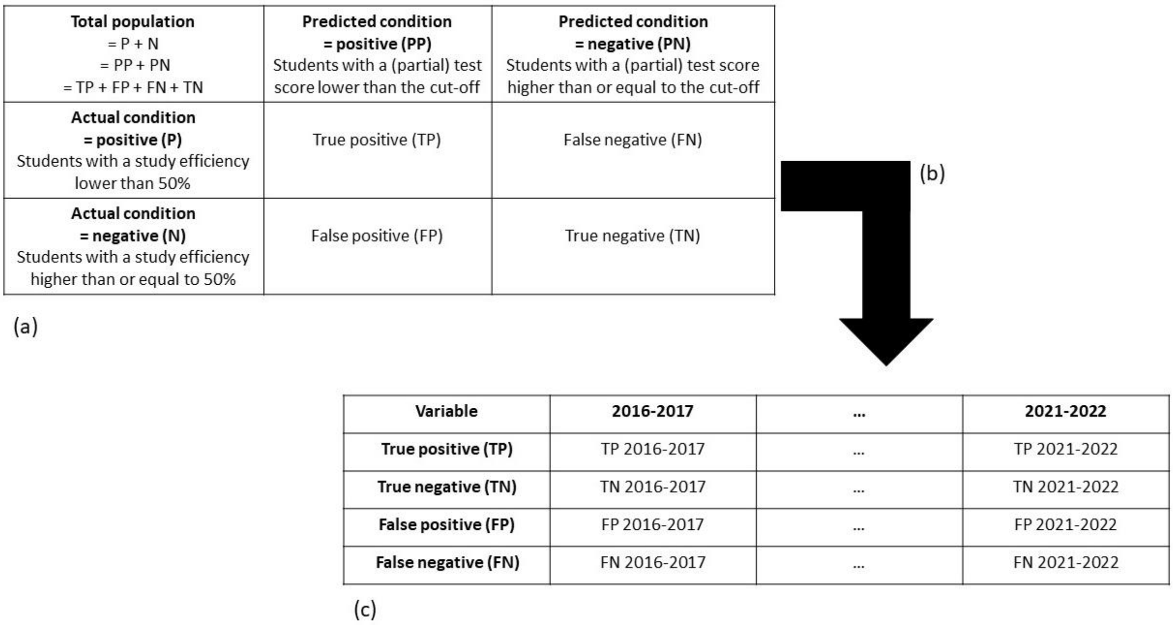
Table 1. Overview of number of participants per test. [1]Test online due to Covid-19. [2]Participation obligatory.

	Study programme	
Year	**ET**	**CBBGG**
2016-2017	52	27
2017-2018	51	25
2018-2019	65	28
2019-2020	115	51
2020-2021	202[1]	47[1]
2021-2022	556[2]	39

First, a comparison of means of study efficiency after the first academic year (amount of ects credits successfully obtained divided by amount of ects credits the student enrolled for, expressed as a percentage), total test scores and partial test scores on standard mathematics and text was performed over the six academic years. Given non-normality of the data and small sample sizes in some cases, non-parametric Kruskal Wallis tests (Kruskal and Wallis 1952, 583-621) and post-hoc Wilcoxon (Wilcoxon, 80-83) tests with sequential Bonferroni-Holm correction for multiple comparison (Holm 1979, 65-70) were used.

For classification purposes, the pseudo-continuous variable study efficiency was categorized into two categories: (i) lower than 50 % and (ii) higher than or equal to 50 %. Given the aim of positioning tests to identify *at-risk* students (i.e. low achievers), the former category was regarded as ‘positive’. Categorized study efficiency was used as dependent variable. Independent variables used were total and partial test scores. A classification was performed for each independent variable and each academic year separately, and for multiple cut-off scores for the independent variables (see Figure 1a). Cut-off scores used were 7, 10, 12 and 14 for total score; 10 for partial score on standard mathematics; 5 on partial score on text. Afterwards, the classifications were repeated with percentile cut-offs of 20 %, 40 %, 60 % and 80 % for total score and 50 % for the partial scores.

Finally, in order compare the stability over the years of the classifications with score cut-offs (criterion-referencing) and percentile cut-offs (norm-referencing), contingency tables with number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) observations for each academic year were constructed separately for each value of cut-off and each test and test part (see Figure 1c) .These contingency tables were subjected to Pearson’s chi-squared test to determine whether statistical differences between the years were present. Given the divergence of means of study efficiency and (partial) positioning test scores from 2020-2021 onwards, the analysis for classification based on total score was repeated for the first four years in the dataset.



Total population = P + N = PP + PN = TP + FP + FN + TN	**Predicted condition** **= positive (PP)** Students with a (partial) test score lower than the cut-off	**Predicted condition** **= negative (PN)** Students with a (partial) test score higher than or equal to the cut-off
Actual condition **= positive (P)** Students with a study efficiency lower than 50%	True positive (TP)	False negative (FN)
Actual condition **= negative (N)** Students with a study efficiency higher than or equal to 50%	False positive (FP)	True negative (TN)

(a)

Variable	2016-2017	...	2021-2022
True positive (TP)	TP 2016-2017	...	TP 2021-2022
True negative (TN)	TN 2016-2017	...	TN 2021-2022
False positive (FP)	FP 2016-2017	...	FP 2021-2022
False negative (FN)	FN 2016-2017	...	FN 2021-2022

(c)

Fig. 1. a) the confusion matrix based on classifications repeated for each academic year, test, test part, cut-off score and cut-off percentile. b) the formation of a contingency table based on data from the confusion matrices, repeated for each test, test part, cut-off score and cut-off percentile. c) an abridged contingency table (middle columns omitted) used for Pearson’s chi-squared test.

3 RESULTS

Figure 2 shows boxplots and comparison of means over the academic year for ET and CBBGG students. The study efficiency of participating ET students is somewhat elevated in 2020-2021. In 2021-2022, the first year of obligatory participation, study efficiency drops again. Note that the study efficiency reported is that only of participants in the positioning test. Oppositely, the CBBGG study efficiency has a decline in 2020-2021. In terms of total test scores and standard mathematics partial scores, an increase can be observed for both ET and CBBGG students in 2020-2021, when the test was held online instead of on campus. Finally, despite the content not changing, a declining trend in partial text score can be observed for ET and CBBGG students over the four academic years with text problems on the test.

Table 2 reports the significance of the Pearson's chi-squared tests for contingency tables of the classifications based on total and partial positioning tests scores of ET and CBBGG students. All total score cut-offs for ET, both criterion-referenced and norm-referenced, yielded significant differences over the years. However, regarding only the first four academic years in the dataset, no norm-referenced cut-off yielded a significant difference. This indicates that the classification into study efficiency groups of ET students using norm-referencing was stable over the first four academic years. For CBBGG students, norm-referencing for total scores was stable over all academic years and both criterion- and norm-referencing were stable over the first four academic years. For ET students, only norm-referencing of partial text scores was stable throughout the academic years. Note that text was only part of the test in the first four academic years, where norm-referencing was also stable for total test scores. For CBBGG students, norm-referencing of standard mathematics and text partial scores were stable, while all criterion-referencing was unstable. Looking at the subset of data before the Covid-19 pandemic, both criterion- and norm-referencing of standard mathematics for both ET and CBBGG students was stable.

4 DISCUSSION AND CONCLUSION

The results show that norm-referenced positioning test score based predictions of academic achievement are generally more stable than criterion-referenced, regardless of any specific cut-off points or test parts. Even for predictions based on the text part, which did not change throughout the four academic years it was used, this conclusion holds. For CBBGG students, norm-referencing was stable throughout the entire six-year-period, despite the extraordinary online edition of the test in 2020-2021 due to the Covid-19 measures and consequent score inflation. For ET students, norm-referencing was only stable up until 2019-2020.

A likely explanation for this difference between CBBGG and ET is that besides Covid-19 measures, participation to the test became obligatory in 2021-2022 (yet results remained non-binding), which changed the composition of the population of participants. It is likely that before the obligation, more motivated students participated on average. It is to be expected that considerable changes in participant population affect the norm-referenced prediction. Another potential explanation for the difference between ET and CBBGG students is that the number of CBBGG participants is lower for each academic year. This means it is harder to find statistically significant evidence for instability which could lead wrong conclusions of stability. While this should be viewed as the most prominent limitation of our study, it does not undermine the conclusion that norm-referencing yields a more stable prediction than criterion-referencing.

Generally, there are considerable advantages of criterion-referencing as well: students deserve a grade that is ‘uncontaminated by reference to how other students in the course perform on the same or equivalent tasks’ (Sadler 2005, 178) and repeated criterion-referencing enables the tracking of progress (Lok, McNaught, and Young 2015, 455). While Lok, McNaught, and Young (2015, 461) state that there is no need for dichotomy between criterion-and norm-referencing, and both can be reported, the fact remains that if decisions for cut-off points need to be made based on historical data, one has to opt for either criterion-or norm referencing. The choice for which mixture of criterion- and norm-referencing is, of course context dependent and specific for each assessment procedure.

The discussion between criterion- and norm-referencing is also relevant for other assessment contexts in higher education. For example, the entrance exam for Medicine and Dentistry in Flanders switched from criterion- to norm-referencing in 2018 because the pass rates were too low before, yearly fluctuations in the number of students are undesirable and it is difficult to keep the difficulty level of the exam problems the same each year (Eggermont 2021, 3). Especially large-scale assessments where the emphasis lies on predictive validity, could benefit from norm-referencing.

In the case of determining cut-offs of positioning tests based on predictions with historical data, the findings of these study recommend using norm-referencing, given the more stable prediction of academic achievement based on norm-referenced positioning test grades, but only when no obvious changes in the population of participants can be expected. This recommendation does, however, not exclude reporting to students their criterion-referenced grade as well. Likely, Flemish students are more used to criterion-referencing, which means that reporting this as well, could increase interpretability of feedback, which is an important issue in the case of positioning tests (Hanssens et al. 2023, 1104). This reflects a specific advantage of reporting criterion-referencing without using it for determining cut-offs.

In conclusion, the decision to use criterion- or norm-referencing depends on the context and goals of the assessment, but reporting both types of grades can be considered to enhance the feedback to students and overall assessment procedure.

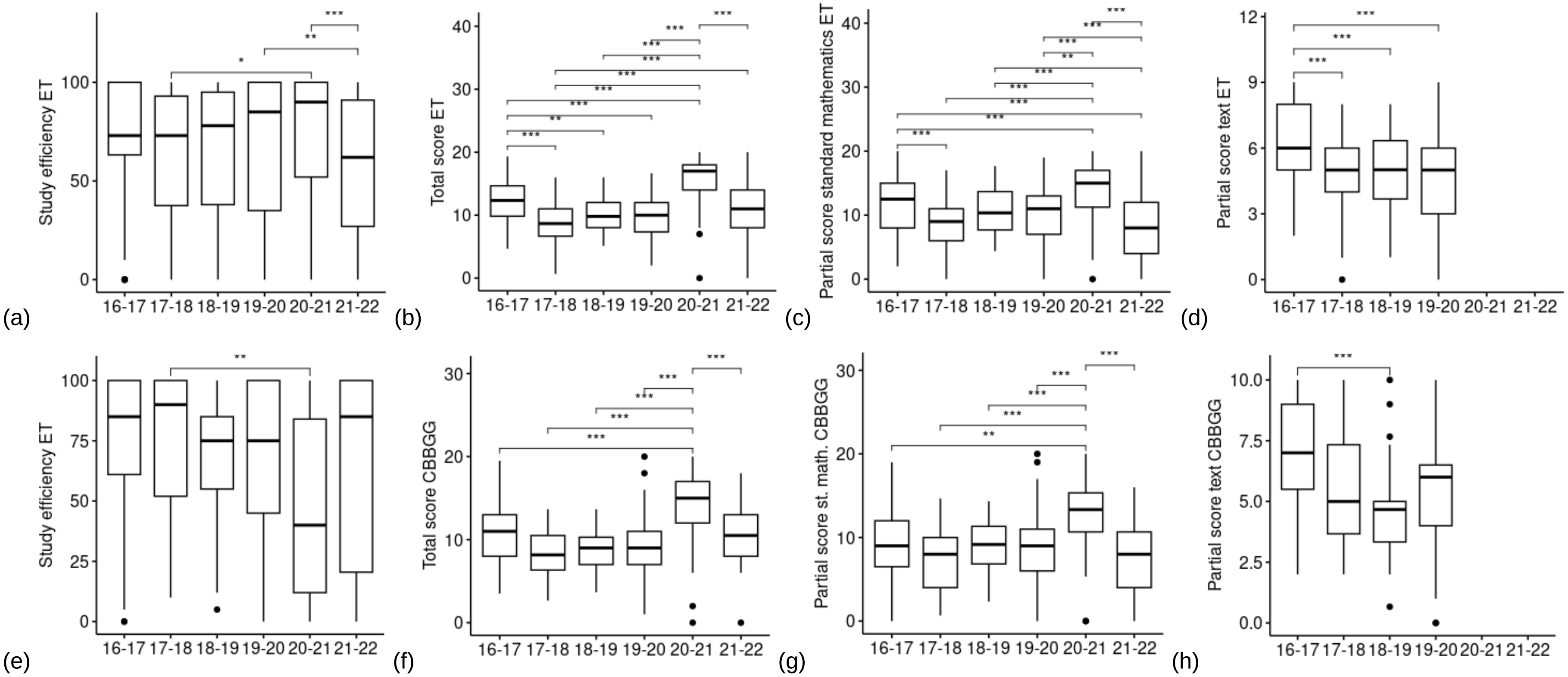


Fig. 2. comparison of means of a) study efficiency, b) total positioning test score, c) partial score on standard mathematics, d) partial score on text for ET students and of e) study efficiency, f) total positioning test score, g) partial score on standard mathematics, h) partial score on text. Significance levels (: 0.05>p>0.01, **: 0.01>p>0.001, ***: p<0.001) of post-hoc Bonferroni-Holm corrected Wilcoxon tests above the corresponding lines. All Kruskal-Wallis tests were significant.*

	Study programme	**Test part**	**Cut-off**	**Chi^2, significance (all academic years)**	**Chi^2, significance (2016-17 – 2019-20)**
Criterion-referenced	ET	Entire test	7	***	n.s. (p=.29)
	ET	Entire test	10	***	**
	ET	Entire test	12	***	***
	ET	Entire test	14	***	n.s. (p=.15)
	ET	Math	10	***	n.s. (p=.17)
	ET	Text	5	/[1]	***
	CBBGG	Entire test	7	**	n.s. (p=.28)
	CBBGG	Entire test	10	***	n.s. (p=.31)
	CBBGG	Entire test	12	***	n.s. (p=.058)
	CBBGG	Entire test	14	**	n.s. (p=.16)
	CBBGG	Math	10	**	n.s. (p=.88)
	CBBGG	Text	5	/[1]	**
Norm-referenced	ET	Entire test	20%	***	n.s. (p=.45)
	ET	Entire test	40%	**	n.s. (p=.96)
	ET	Entire test	60%	***	n.s. (p=.28)
	ET	Entire test	80%	*	n.s. (p=.97)
	ET	Math	50%	**	n.s. (p=.85)
	ET	Text	50%	/[1]	n.s. (p=.63)
	CBBGG	Entire test	20%	n.s. (p=.069)	n.s. (p=.63)
	CBBGG	Entire test	40%	n.s. (p=.31)	n.s. (p=.84)
	CBBGG	Entire test	60%	n.s. (p=.23)	n.s. (p=.49)
	CBBGG	Entire test	80%	n.s. (p=.35)	n.s. (p=.84)
	CBBGG	Math	50%	n.s. (p=.11)	n.s. (p=.93)
	CBBGG	Text	50%	/[1]	n.s. (p=.26)

*Table 2. Significance levels of Pearson's chi-squared tests for contingency tables of classification based on total and partial positioning test scores of ET and CBBGG students. [1]Text was only part of the test in 2016-17 - 2019-2020. (n.s.: p>0.05., *: 0.05>p>0.01, **: 0.01>p>0.001, ***: p<0.001)*

REFERENCES

Eggermont, J. 2021. “Het toelatingsexamen arts anno 2021.” *Periodiek – driemaandelijks tijdschrift van het Vlaams Geneeskundigenverbond* 76: 24-28.

Hanssens, J., Langie, G., and Van Soom, C. 2023. “Students’ perceptions of low stakes positioning tests at the start of higher STEM education: A mixed methods approach.” International Journal of Education in Mathematics, Science, and Technology (IJEMST) 11, no. 5: 1094-1112. https://doi.org/10.46328/ijemst.2889

Holm, S. 1979. “A Simple Sequentially Rejective Multiple Test Procedure.” *Scandinavian Journal of Statistics* 6, no. 2: 65–70.

Kruskal, W. H., and Wallis, W. A. 1952. “Use of Ranks in One-Criterion Variance Analysis.” *Journal of the American Statistical Association* 47, no. 260: 583–621.

Lok, B.C.Y., McNaught, C., and Young, K. 2015. “Criterion-referenced and norm-referenced assessments: compatibility and complementarity.” *Assessment & Evaluation in Higher Education* 41, no. 3: 450-465.

Pinxten, M., Van Soom, C., Peeters, C., De Laet, T., and Langie, G. 2019. “At-risk at the gate: prediction of study success of first-year science and engineering students in an open-admission university in Flanders - Any incremental validity of study strategies?” *European Journal of Psychology of Education* 34, no. 1: 45–66.

Sadler, D.R. 2005. “Interpretations of criteria-based assessment and grading in higher education.” *Assessment & Evaluation in Higher Education* 30, no. 2: 175-194.

Vanderoost, J., Callens, R., Vandewalle, J., and De Laet, T. 2014. “Engineering positioning test in Flanders : a powerful predictor for study success?” *Proceedings of the 42nd Annual SEFI Conference.*

Vanderoost, J., Van Soom, C., Langie, G., Van Den Bossche, J., Callens, R., Vandewalle, J., and De Laet, T. 2015. “Engineering and science positioning tests in Flanders: Powerful predictors for study success?” *Proceedings of the 43rd SEFI Annual Conference.*

Vandewalle, J.P.L., and Callens, R. 2013. “A positioning test mathematics in Flanders for potential academic engineering students.” *41st SEFI conference.*

Van den Broeck, L., De Laet, T., Lacante, M., Pinxten, M., Van Soom, C., and Langie, G. 2019. “Predicting the academic achievement of students bridging to engineering: the role of academic background variables and diagnostic testing.” *Journal of Further and Higher Education* 43, no. 7: 989–1007.

Wilcoxon, F. 1945. “Individual Comparisons by Ranking Methods.” *Biometrics Bulletin* 1, no. 6: 80–83.