112-1 Statistical Methods
Final Report
學習表現進步的關鍵變數

數據所 RE6124035 黃亮臻

January, 2024

壹、 動機與目標

求學生涯中，人們不乏追求好的成績，過去也有多篇研究探討影響學業成績的關鍵變數。儘管成績的優異無疑是值得追求的，但對於在學習過程中取得進步的學生，他們的努力與成長同樣值得肯定。因此有了本次的研究動機：透過建立迴歸模型，探討哪些變數對學習表現的提升有顯著的影響。

貳、 預期困難

一、 何謂進步

為了探究學生學習進步的情況，我以成績作為衡量學習成效的指標。鑑於資料集涵蓋的兩次考試範圍不同，其難度可能各異，若直接比較兩次分數的差異可能有失準確性。為處理此問題，我使用兩種不同的資料預處理方法進行調整，將於第四節的資料說明做介紹。透過這些調整，雖然建立了一個更可靠的衡量進步的基準，但也同時使得後續解釋變得較不直觀。

二、 資料型態

此份資料大多為 ordinal 與 nominal 的變數，只有少數為連續型變數，且這些連續型變數呈現出相當分散的分佈特性。這樣的結構可能會為建立模型帶來一定的挑戰。

參、 文獻探討

國外學者 Camilla Molin 在 2020 年使用同一資料集進行了研究，以「是否被當（成績小於 10）」作為應變數，使用 Logistic regression 進行分析。除此之外，該學者還對葡萄牙的教育系統進行了詳細介紹，並對資料集中包含的兩所學校提供了深入的描述。雖然此份研究與本次的研究方向不同，但他對於變數之間關聯性的分析成果，仍有參考的價值。

肆、 資料說明

此份資料包含兩份資料集，第一份是 395 位學生的數學成績以及其父母學歷、職業以及學生的生活規劃等自變數，第二份 649 位學生的葡萄牙語成績以及其父母學歷、職業以及學生的生活規劃等自變數。本次分析將以數學成績的資料集為主。

一、 應變數

本次分析的目標為探討哪些因素對學生的學習表現提升有顯著影響。資料集中包含第一次及第二次考試成績，鑑於兩次試卷的難易度可能不同，為了消去不同試卷難度帶來的影響，我採用以下兩種方法進行調整：

1. 兩次成績去中心化後的分數差異

從學生在兩次考試得到的成績中，分別減去對應考試的平均成績，從而計算出每位學生的成績相對於班級平均水平的高低。進一步地，將兩次考試的去中心化成績差值相減，得出的數值用來衡量學生的學習進步幅度。若差值為正，表示學生在學習上有進步；若為負，則表示學習上有所退步。

2. 兩次考試的排名差異

使用學生在班級中的排名，量化學生的學習表現。通過計算學生在第一次和第二次考試中排名的變化來衡量他們的學習進步。此應變數為第二次考試排名進步的多寡，因此若數值為負，表示學生表現進步；反之則為退步。

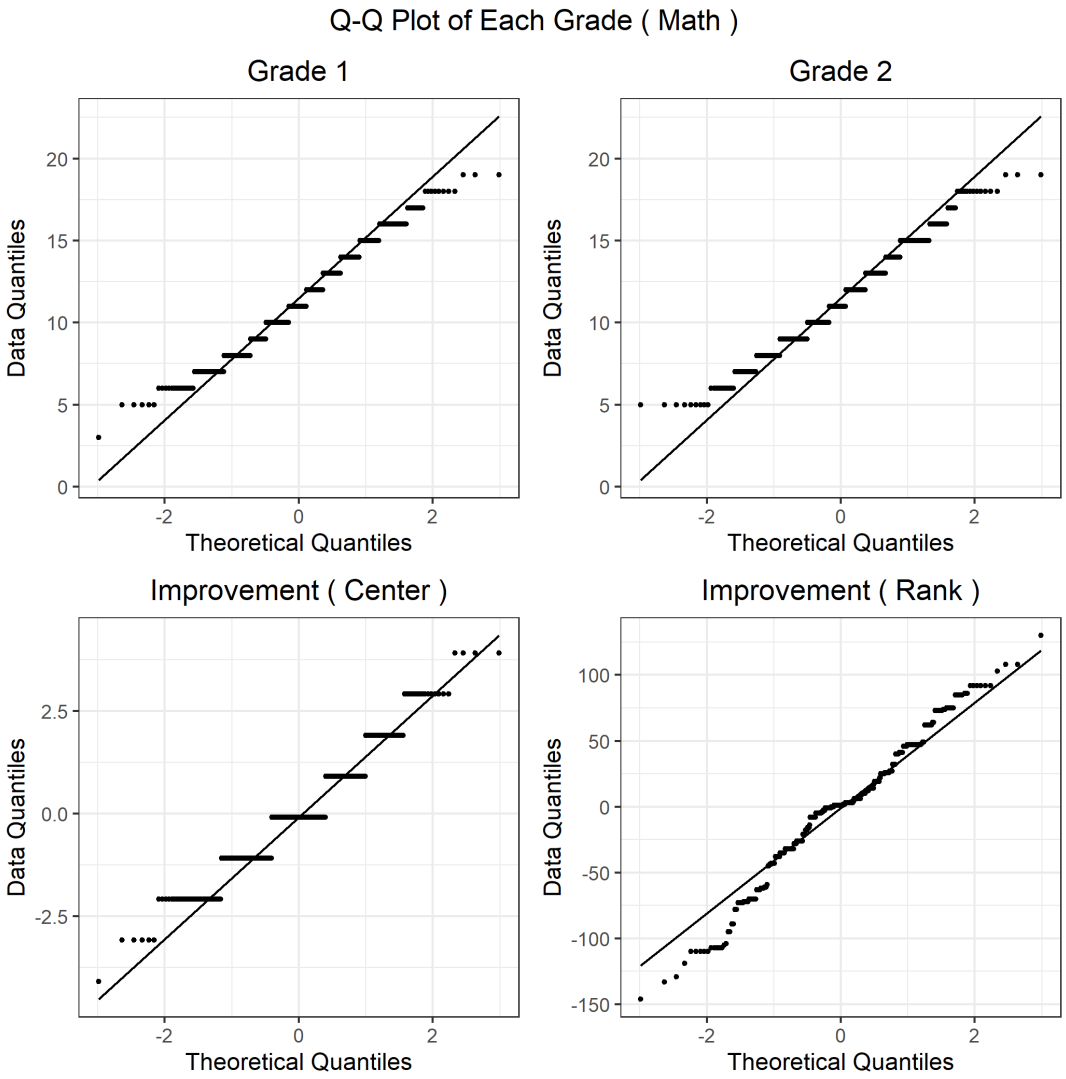

圖 1：第一次、第二次成績與兩種應變數的 Q-Q Plot

二、 自變數

此份資料集共有 30 個自變數，且均無遺失值。各變數的說明如下：

變數名稱	資料型態	說明
school	binary	就讀學校（"GP"：Gabriel Pereira、"MS"：Mousinho da Silveira）
sex	binary	性別（"F"：女性、"M"：男性）
age	numeric	年齡
address_type	binary	居住位置（"Urban"、"Rural"）
family_size	ordinal	家庭大小（"≤3"、">3"）
parent_status	binary	居住狀況（"Living together"、"Apart"）
mother_education	ordinal	母親教育程度（"none"、"primary(4th grade)"、 "5th-9th grade"、"secondary"、 "higher"）
father_education	ordinal	父親教育程度（"none"、"primary(4th grade)"、 "5th-9th grade"、"secondary"、 "higher"）
mother_job	nominal	母親職業（"teacher"、"health"：醫療保健相關、"services"：公務員、"at_home"、"other"）
father_job	nominal	父親職業（"teacher"、"health"：醫療保健相關、"services"：公務員、"at_home"、"other"）
reason	nominal	選擇學校原因（"home"：離家近、"reputation" ：學校聲譽、"course"：課程偏好、"other"）
guardian	nominal	主要照顧者（"mother"、"father"、"other"）
travel_time	ordinal	到校通勤時間（"<15"、"15-30"、"30-60"、 ">60"；單位：分鐘）
study_time	ordinal	每週學習時間（"<2"、"2-5"、"5-10"、">10"；單位：小時）
class_failures	ordinal	過往被當次數（n if 1<=n<3；else 4）
school_support	binary	額外教育支持（"yes"、"no"）
family_support	binary	家庭教育支持（"yes"、"no"）
extra_paid_classes	binary	額外付費課程（"yes"、"no"）
activities	binary	額外課外活動（"yes"、"no"）
nursery	binary	參加安親班（"yes"、"no"）

變數名稱	資料型態	說明
higher_ed	binary	希望得到更高教育（"yes"、"no"）
internet	binary	在家可否使用網路（"yes"、"no"）
romantic_relationship	binary	有戀愛關係（"yes"、"no"）
family_relationship	ordinal	家庭關係（1：非常糟糕－5：非常良好）
free_time	ordinal	課後自由時間（1：非常少－5：非常多）
social	ordinal	與同儕外出時間（1：非常少－5：非常多）
weekday_alcohol	ordinal	平日飲酒量（1：非常少－5：非常多）
weekend_alcohol	ordinal	假日飲酒量（1：非常少－5：非常多）
health	ordinal	健康狀況（1：非常糟糕－5：非常良好）
absences	numeric	缺課次數

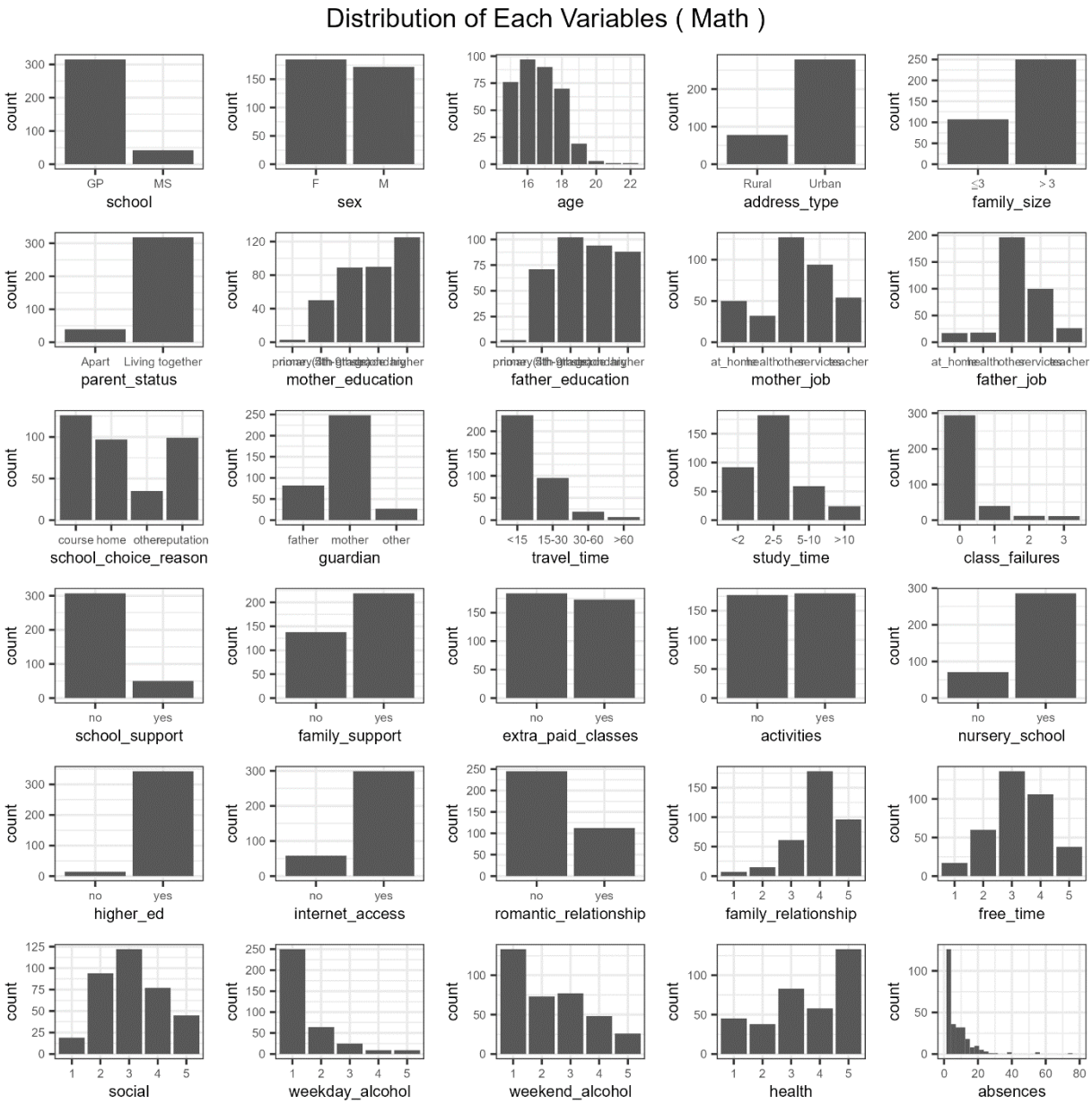

圖 2：自變數分布狀況

伍、 分析方法

在本研究中，我使用全部的自變數分別對兩個應變數建立迴歸模型，並從 30 個自變數中，篩選出 p-value 小於 0.1 的變數，再進一步簡化迴歸模型。總共建立了四個模型，如下表 1。

模型 1 以去中心化資料作為應變數，納入所有的自變數；模型 2 則是基於模型 1，只包含了 p-value 小於 0.1 的自變數；模型 3 以排名進步資料作為應變數，納入所有自變數；模型 4 則從模型 3 中篩選出 p-value 小於 0.1 的自變數。透過四個模型的比較與分析，希望能找出更為精簡且解釋力更強的迴歸模型。

表 1：四種模型其自變數的 p-value

變數名稱	model 1	model 2	model 3	model 4
school	0.0188	0.0159	0.0552	0.0488
sex	0.9678		0.8229	
age	0.0001	<0.0001	<0.0001	<0.0001
address_type	0.7359		0.6727	
family_size	0.9058		0.9896	
parent_status	0.9977		0.7413	
mother_education	0.5415		0.5294	
father_education	0.5254		0.7192	
mother_job	0.2737		0.1640	
father_job	0.5773		0.5717	
reason	0.0522	0.1542	0.2105	
guardian	0.9751		0.8580	
travel_time	0.9810		0.8667	
study_time	0.9452		0.9449	
class_failures	0.0240	0.0206	0.0813	0.1242
school_support	0.5305		0.7482	
family_support	0.2146		0.4211	
extra_paid_classes	0.2226		0.2286	
activities	0.8953		0.7984	
nursery	0.9942		0.7819	
higher_ed	0.2078		0.5565	
internet	0.1333		0.1085	
romantic_relationship	0.8953		0.7368	
family_relationship	0.9075		0.7310	
free_time	0.0103	0.0025	0.0110	0.0031
social	0.5090		0.3875	
weekday_alcohol	0.8580		0.8443	
weekend_alcohol	0.1135		.2003	

變數名稱	model 1	model 2	model 3	model 4
health	0.4729		0.3732	
absences	0.0087	0.0398	0.0123	0.0278
Model adjusted R2	0.0991	0.1437	0.0822	0.1292
Model p-value	0.0211	<0.0001	0.0441	<0.0001

本研究中所有模型的 Adjusted R-squared 普遍低於 0.15，表示模型未能充分解釋資料的變異。以下為對於 Adjusted R-squared 相對較高的模型 2 和模型 4，進行迴歸假設的檢查，並檢視離群值是否對模型結果造成影響。

一、 樣本間相互獨立

使用 Durbin-Watson 檢定殘差項是否存在自我相關。由下表 3 得知，D-W 值接近 2，p-value 均大於 0.05，傾向接受 H_0：誤差項無自我相關，推測滿足獨立性假設。

表 3：模型 2 與模型 4 之獨立性檢定

	imp_center model 2	imp_rank model 4
D-W Statistic	2.121	2.100
P-value	0.298	0.408

二、 常態性假設

使用 Shapiro-Wilk 檢定殘差是否為常態。由下表 4 得知，p-value 均大於 0.05，傾向接受 H_0：殘差為常態分佈。

表 4：模型 2 與模型 4 之常態性檢定

	imp_center model 2	imp_rank model 4
W Statistic	0.9918	0.9934
P-value	0.1163	0.2446

同時，繪製兩模型殘差的 Q-Q plot，如下圖 3。大部分殘差都很接近 45 度線，只有頭尾兩端稍微偏離。因此，不論是 Shapiro-Wilk 檢定或是 Q-Q plot 的呈現，都可以推測資料滿足常態性假設。

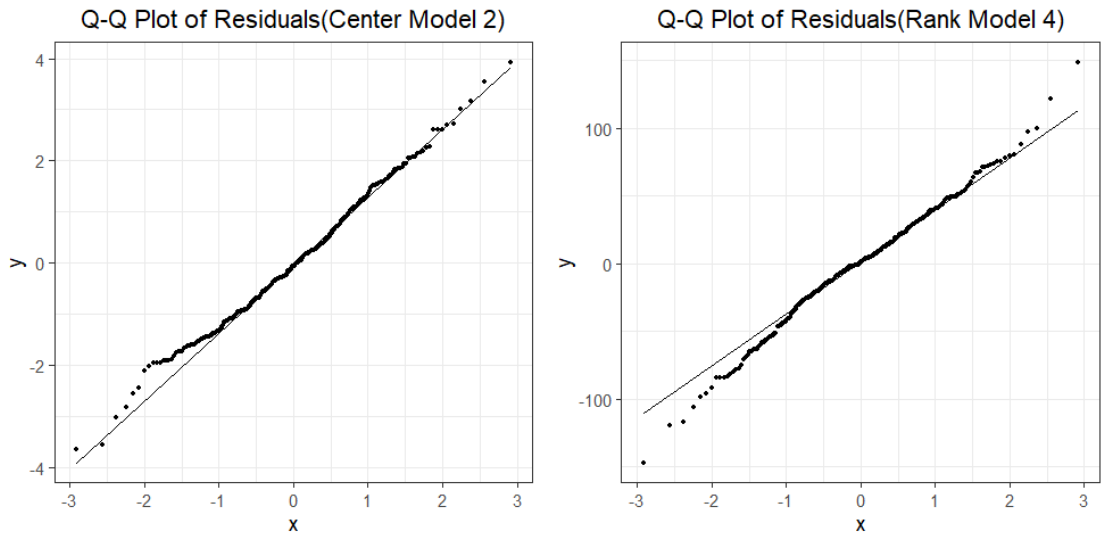

圖 3：模型 2 與模型 4 的殘差 Q-Q plot

三、 變異數同質性

整體而言兩個模型的殘差並沒有明顯的趨勢，均以 0 為中心上下分散，因此推測滿足變異數相等的假設。

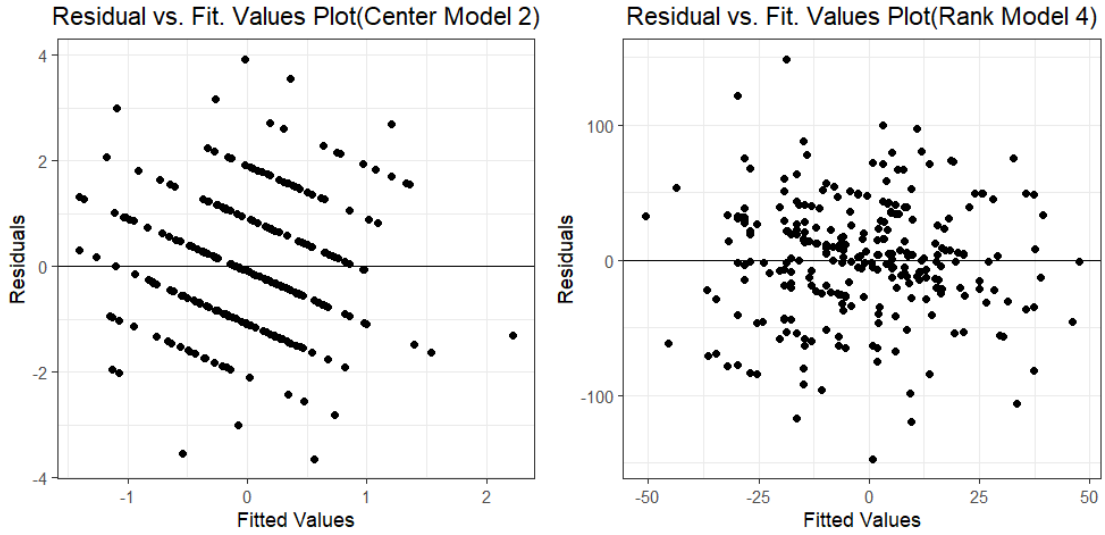

圖 4：模型 2 與模型 4 的殘差分佈

四、 是否受離群值影響

由下表 2 得知，Cook's Distance 在模型 2 與模型 4 中最大值均不超過 0.1，因此判斷模型並沒有受離群值的影響。

表 2：模型 2 與模型 4 之 Cook's Distance 平均值與最大值

	model 2	model 4
平均值	0.0037	0.0038
最大值	0.0630	0.0558

陸、 研究結果

本次研究的兩個模型 Adjusted R-squared 皆偏低，去中心化的模型（Model 2） 其 Adjusted R-squared 為 0.14，排名的模型（Model 4）則為 0.13。值得注意的是，儘管去中心化的應變數顯示出間斷性（見圖 1），班級排名的應變數看似更加連續，但去中心化資料的模型展現了略高的解釋力，與預想的狀況稍微不同。

僅管兩個模型仍有很多變異沒有被解釋，但考量社會研究的模型解釋力經常不高，且經過檢定各項假設均無違背，因此認為本次研究的兩個模型仍可以爲我們提供一些發現。模型的具體係數詳見表 5 及表 6。

表 5：去中心化資料（模型 2）的各項係數

變數	係數	p-value
(Intercept)	5.235	<0.001
schoolMS	-0.194	0.473
age	-0.311	<0.001
school_choice_reason home	-0.101	0.613
school_choice_reason other	0.473	0.090
school_choice_reason reputation	-0.121	0.548
class_failures	0.364	0.002
free_time.L	-0.726	0.010
free_time.Q	-0.176	0.466
free_time.C	-0.546	0.007
free_time^4	0.271	0.076
absences	-0.02	0.039

表 6：班級排名資料（模型 4）的各項係數

變數	係數	p-value
(Intercept)	-178.821	<0.001
schoolMS	-0.266	0.976
age	10.595	<0.001
class_failures	-9.357	0.022
free_time.L	25.161	0.007
free_time.Q	7.815	0.337
free_time.C	15.915	0.018
free_time^4	-8.007	0.118
absences	0.712	0.027

由上表可發現，兩種模型的係數正負值可能不同，這種差異來自於兩種不同的應變數衡量方式。去中心化後的分數差異模型 (Model 2) 中，學習進步與係數為正向的關係；而在班級排名模型 (Model 4) 中，進步與係數為反向的關係。另外，儘管兩種模型的應變數都旨在衡量學習進步，但他們篩選出來的自變數卻有些不同。對於兩種模型都有顯著影響的變數如下：年齡、

過往被當次數、自由時間以及缺課次數。分別解釋為：年幼學生相較年長學生進步更為明顯；過往被當次數越多的學生進步較不明顯；自由時間較少的學生往往進步更快；缺課次數少的學生傾向有更大的進步幅度。

柒、 結論與建議

本次研究透過兩種方法定義學習進步，找出了影響學生學習表現進步的關鍵變數。然而，由於模型的 Adjusted R-squared 值偏低，可能表示還有其他未考慮到的關鍵因素。為了更深入的理解學習表現進步背後的原因，未來的研究可以嘗試探索其他潛在因素，並將其納入研究範疇中。

捌、 參考文獻

1. Camilla Molin (2020). A statistical analysis of the performance in mathematics of secondary student in Portugal. U.U.D.M Project Report 2020:28
2. Lulu Cheng (2017). Exploring the Factors that Affect Secondary Student's Mathematics and Portuguese Performance in Portugal. Technological University Dublin