



How to achieve quality in medical education? Inter-rater agreement about item-writing flaws in multiple-choice questions: the case of Clinical Anatomy

B. Guimarães^{1,2,3,4}, J. Pais², E. Coelho², A. Povo^{1,2,3,5}, J. Ribeiro^{1,2}, B. Cruz^{1,2}, A. Ferreira^{1,2,3}, S. Tsisar^{1,3}, L. Dourado^{1,3}, C. Carvalho², M.A. Ferreira^{1,2}

¹ Department of Anatomy, Faculty of Medicine of the University of Porto, PORTUGAL

² Department of Medical Education and Simulation, Faculty of Medicine of the University of Porto, PORTUGAL

³ CINTESIS - Center for Research in Health Technologies and Information Systems. Faculty of Medicine, University of Porto, PORTUGAL

⁴ General Resident, Centro Hospitalar de Entre o Douro e Vouga, Santa Maria da Feira, PORTUGAL

⁵ General Surgery, Centro Hospitalar do Porto, Porto, PORTUGAL

Porto

V Simpósio AMP

4 de novembro de 2016

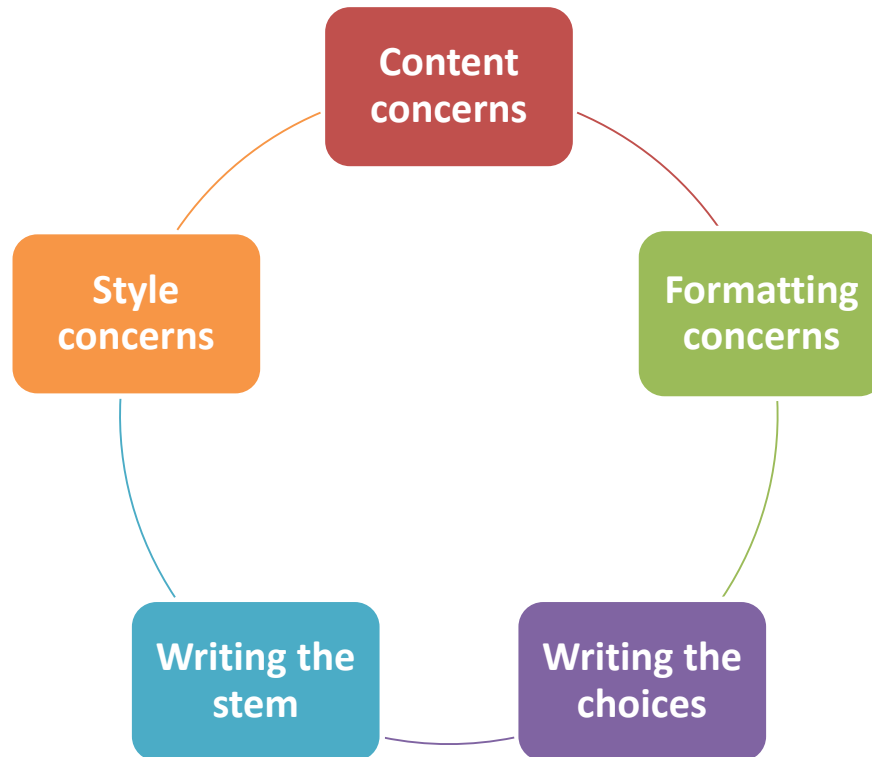
Secção Regional Norte da Ordem dos Médicos

Assessment in Medicine...

Multiple-choice questions (MCQs) are frequently used to assess students in health science disciplines.

The MCQ format allows teachers to efficiently assess large number of students and also to test a wide range of content.

Haladyna and his collaborators (2002) displayed 31 principles in a review article. These principles are divided in five areas, and they summarized current educational measurement recommendations concerning MCQ item writing.



MCQs`Rules Code:

	RuleID	Rule
CONTENT CONCERNS	0	No problem
	1	Single content and behavior
	2	Important, not trivial content
	3	3Use novel material
	4	Keep items independent
	5	Avoid over specific/general
	6	Avoid opinions
	7	Avoid trick items
FORMATTING CONCERNS	8	Simple vocabulary
	9	Avoid the complex (typeK) format
STYLE CONCERNS	10	Format vertically
	11	Edit and proof
	12	Correct grammar
	13	Minimize readingh
WRITING THE STEM	14	Clear directions
	15	Central idea in stem
	16	Avoid window dressing
	17	Use positive, no negatives
	18	Write as many plausible distractors as you can
WRITING THE CHOICES	19	One right answer
	20	Vary location of right answer
	21	Logical/numerical order
	22	Choices not overlapping
	23	Choices homogeneous
	24	Choice length equal
	25	Use carefully None of the above
	26	Avoid All of the above
	27	Avoid NOT in choices
	28	Avoid clues
	29	Make distractors plausible
	30	Use common errors of students
	31	Use humor sparingly

The aim of this study was to estimate the inter-rater agreement on item classification as either standard or flawed

**1st Step:
Judgers**

- Four judges (2 teacher/2 students), blinded to all item performance data, independently classified each item.

**2nd Step:
MCQ Items**

- 920 test items from 10 examinations (2008-2012) were divided by field of knowledge and were classified as either standard or flawed.

**3rd Step:
Proceeding**

- If flawed the exact type of item flaw or flaws contained within the question (including options) was recorded.

**4th Step:
Control Item**

- A standard item was operationally defined for this study as any item that did not violate one or more of the 31 principles.

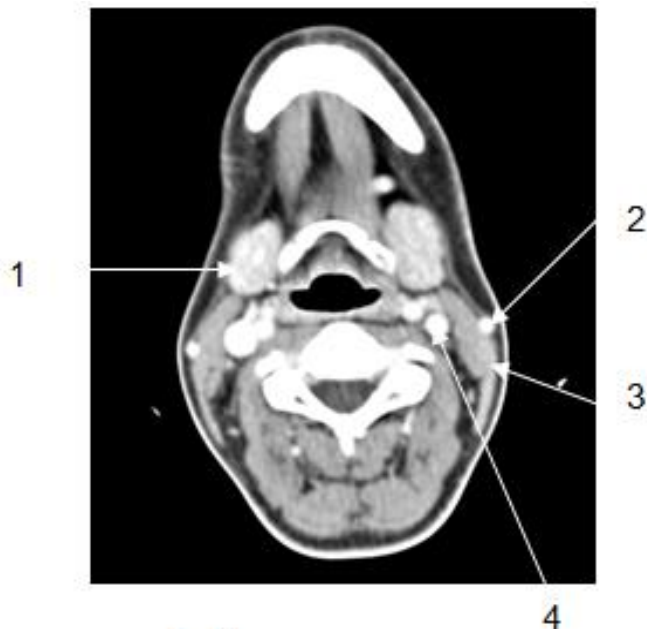
Statistical Analysis

The Fleiss' Kappa was used to evaluate the inter-rater agreement between 4 judges previous to the consensus process.

Guidelines for interpreting kappa statistics :

- 0.81 - 1.00 indicate almost perfect agreement
- 0.61 - 0.80 substantial agreement
- 0.41 - 0.60 moderate agreement
- 0.21 - 0.40 fair agreement
- less than 0.21 are poor or slight agreement.

A 35 years old man comes to your office with complains of cervical pain during the meals. He is submitted to a TC scan. **Signalize the wrong option:**



- a) "1" is related to the facial vases
- b) On the phisical exam it is possible to detect a small swelling on the stuture "4" during the Valsalva's Manoeuvre (counter resistance forced expriration)
- c) "2" crosses the posterior triangle of the neck
- d) "3" is responsable for the contralateral rotation of the neck
- e) "1" is independent of the deep cervical fascia

17. Use positive, no negatives

24. Choice length equal

		N(%) ¹	Inter-rater Reliability ²	Agreement by at least 3 raters N (%) ³
1	Abdomen	709 (19.3)	0.88	167 (16.2)
2	Head	380 (10.3)	0.80	88 (9.6)
3	Lower Limb	336 (9.1)	0.94	81 (8.8)
4	Neck	577 (15.7)	0.82	131 (14.2)
5	Pelvis and Perineum	647 (17.5)	0.94	160 (17.4)
6	Thorax	659 (17.9)	0.89	158 (17.2)
7	Upper Limb	338 (9.2)	0.95	83 (9.0)
	Absolute agreement			868 (94.3)

¹From a total 920 items multiplied by 4 raters

²Fleiss kappa

³From a total 920 itens

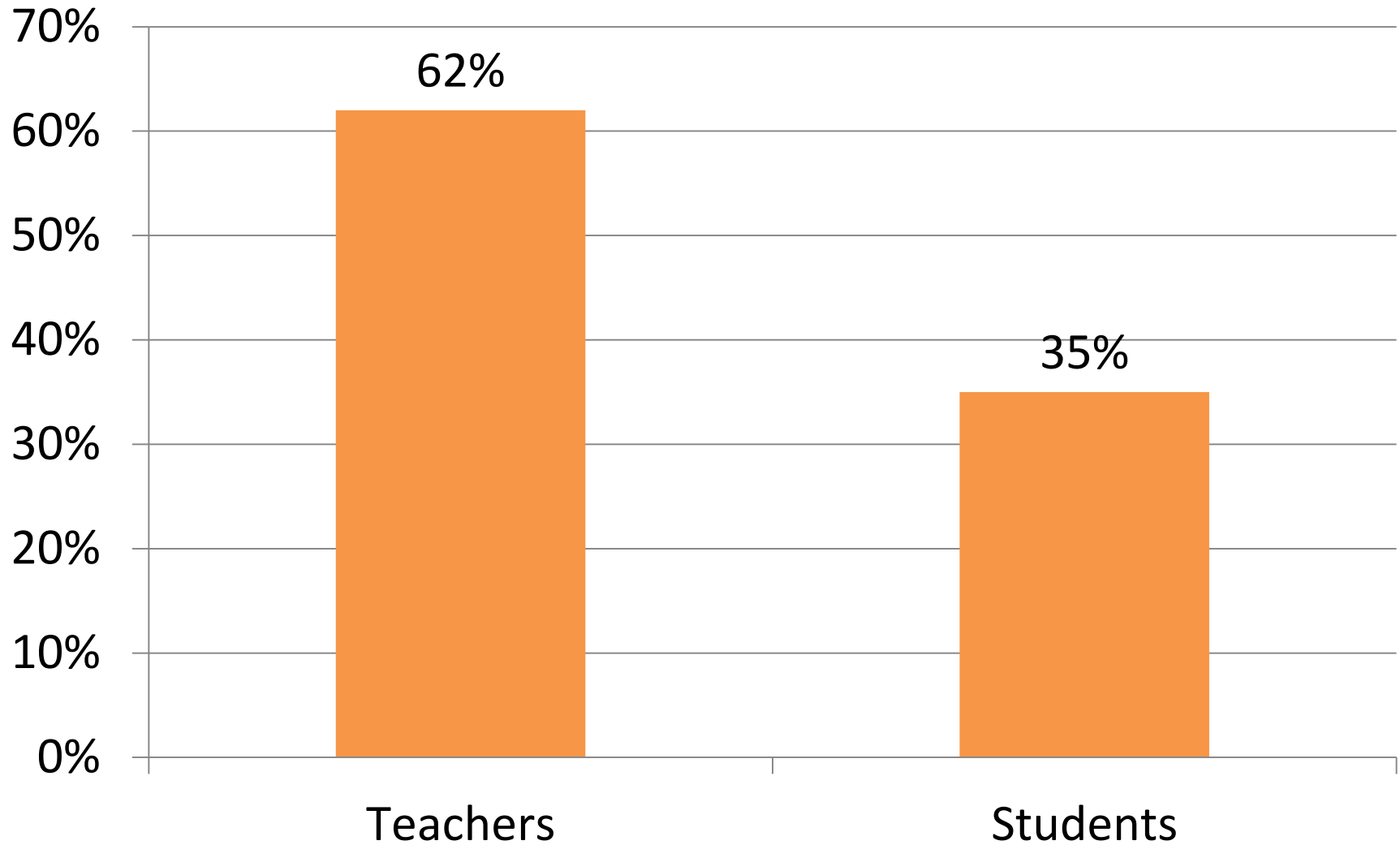


Figure 1 - Prevalence of Flaw MCQ Items by Teachers Group and Students Group

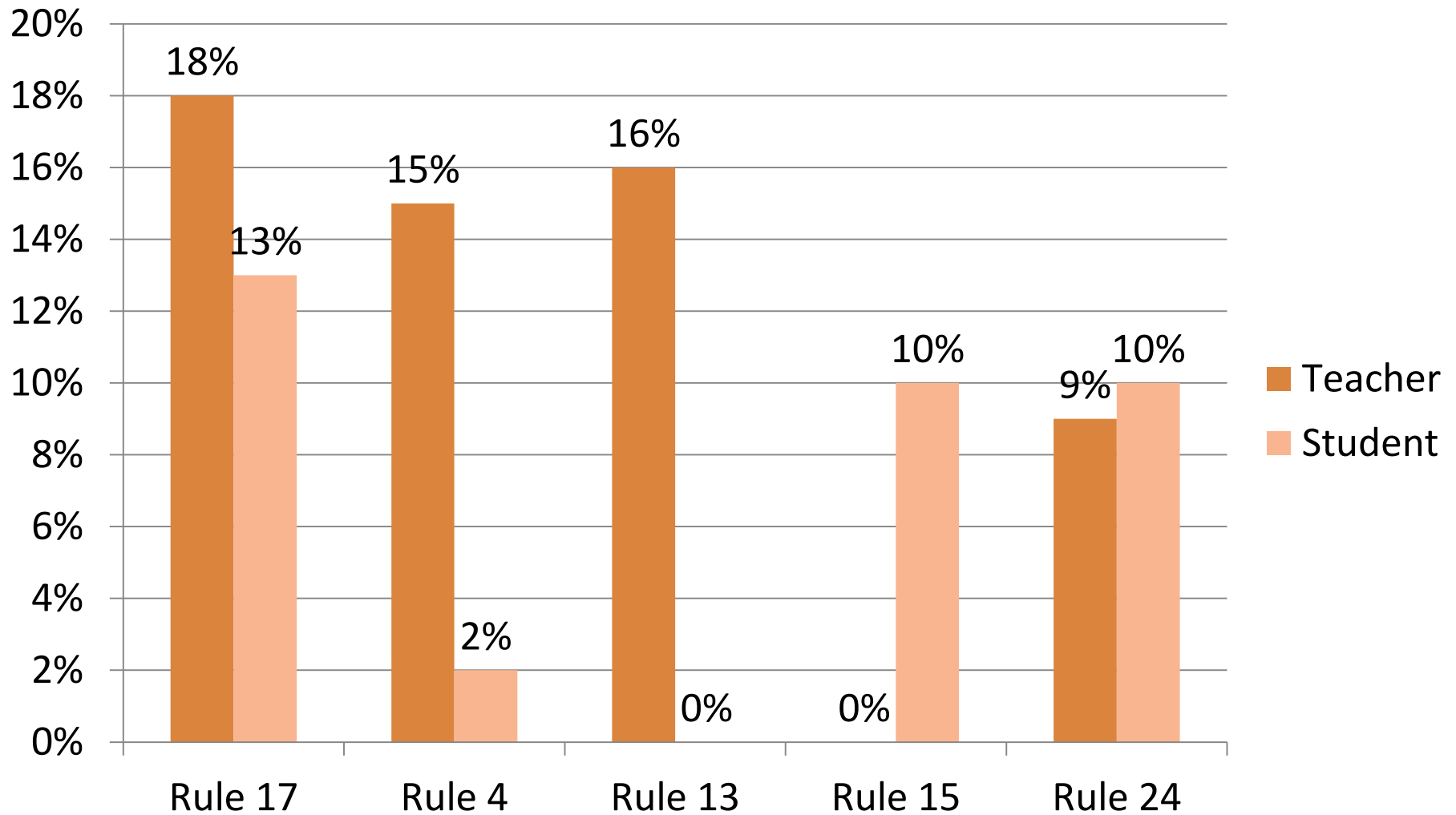


Figure 2 - Prevalence of flaws by rule between teachers and students

Background	Participants	Methods	Results	Discussion	Take Home Message
------------	--------------	---------	---------	------------	-------------------

Principle ID	Rule	N	(%) ¹	Inter-rater Reliability ²	Agreement by at least 3 raters N (%) ³
0	No Problem	1903	51.7	0.3	374 (40.7)
17	Use positive. No negatives	566	15.4	0.7	155 (16.8)
24	Choice length equal	349	9.5	0.5	53 (5.8)
4	Keep items independent	309	8.4	0.3	28 (3.0)
13	Minimize reading	298	8.1	-0.1	0 (0.0)
15	Central idea in stem	196	5.3	0.0	0 (0.0)
25	Use carefully none of the above	167	4.5	0.8	45 (4.9)
22	Choices not overlapping	126	3.4	0.1	3 (0.3)
26	Avoid all of the above	62	1.7	0.8	14 (1.5)
10	Format vertically	56	1.5	0.0	0 (0.0)
27	Avoid NOT in choices	54	1.5	0.1	0 (0.0)

¹From a total 920 items multiplied by 4 raters

²Fleiss kappa

³From a total 920 itens

Principle ID	Rule	N	(%) ¹	Inter-rater Reliability ²	Agreement by at least 3 raters N (%) ³
12	Correct grammar	46	1.3	0.0	0 (0.0)
28	Avoid clues	42	1.1	0.0	0 (0.0)
8	Simple vocabulary	39	1.1	0.0	0 (0.0)
1	Singe content and behaviour	38	1.0	0.0	0 (0.0)
16	Avoid window dressing	33	0.9	0.0	0 (0.0)
2	Important. Not trivial content	32	0.9	0.0	0 (0.0)
23	Choices homogeneous	29	0.8	0.0	0 (0.0)
14	Clear directions	25	0.7	0.1	0 (0.0)
19	One right answer	19	0.5	0.2	1 (0.1)
7	Avoid trick items	15	0.4	0.0	0 (0.0)

¹From a total 920 items multiplied by 4 raters

²Fleiss kappa

³From a total 920 itens

Principle ID	Rule	N	(%) ¹	Inter-rater Reliability ²	Agreement by at least 3 raters N (%) ³
29	Make distractors plausible	8	0.2	0.0	0 (0.0)
5	Avoid over specific/general	5	0.1	0.0	0 (0.0)
6	Avoid opinions	5	0.1	0.0	0 (0.0)
31	Use humour sparingly	5	0.1	0.0	0 (0.0)
21	Logical/numerical order	3	0.1	0.0	0 (0.0)
3	Use novel material	0	0.0	---	---
18	Write as many plausible distractors as you can	0	0.0	---	---
20	Vary location of right answer	0	0.0	---	---
30	Use common errors of students	0	0.0	---	---
9	Use the question, completion...	-	-	-	-

¹From a total 920 items multiplied by 4 raters

²Fleiss kappa

³From a total 920 items

- The prevalence of flaws detected by the Teacher group is higher in comparison with the one of the Students group
- Even the rules more selected are different, the most prevalent rules in Teacher Group are 17, 13 and 4 and for students are 17, 24 and 15.
- The results showed many disagreements among the 4 judges regarding item classification, previous to the consensus process; however for the most prevalent principles the agreement was substantial;

- In fact, the agreement was moderate/substantial for the following principles:
 - 17. Use positive, no negatives;
 - 24. Choice length equal;
 - 25. Use carefully “none of the above”;
 - 26. Avoid “all of the above”
- The rules that show moderate to substantial agreement are most objective.
- Generally the rules that showed poor or slight agreement are more subjective. For instance, rule 13. *Minimize reading*.

- Even without the consensus reunion the preliminary prevalence of item flaws is high
- The absolute agreement about item's field division was almost perfect agreement
- In order to clarify the principles among the rater it must be necessary formation and an extensive review of the literature.
- In the future, it is important to measure the impact that this flaws have on the quality indicator of a examination (difficulty indices and discrimination indices), and by this way compare with the existing results

It is important to increase the formal instruction in MCQs item-writing in order to minimize item-writing flaws in future examinations.

V Simpósio AMP

4 de novembro de 2016

Secção Regional Norte da Ordem dos Médicos

תודה
 Dankie Gracias
 Спасибо شكراً
 Merci Takk
 Köszönjük Terima kasih
 Grazie Dziękujemy Děkojame
 Ďakujeme Vielen Dank Paldies
 Kiitos Täname teid 谢谢
Thank You Tak
 感謝您 Obrigado Teşekkür Ederiz
 Σας ευχαριστούμε 감사합니다
 Bedankt Дěkujeme vám
 ありがとうございます
Tack

Bruno Guimarães

btsguimaraes@med.up.pt / btsguimaraes@gmail.com

t: +351 92 778 49 01/ +351 91 191 90 23

Department of Anatomy, Faculty of Medicine of the University of Porto
 Department of Medical Education and Simulation, Faculty of Medicine of
 the University of Porto

CINTESIS - Center for Research in Health Technologies and Information
 Systems. Faculty of Medicine, University of Porto

General Resident, Centro Hospitalar de Entre o Douro e Vouga, Santa
 Maria da Feira