Resultados das Provas de Aferição e dos Exames de Matemática e de Português do 9º ano: Que conclusões tirar?

Carlos Pereira dos Santos[1], Luísa Araújo[2] & Margarida Graça[3]

O presente artigo resulta de uma análise crítica sobre os resultados das Provas de Aferição e dos Exames de Matemática e de Português do 9º ano do Ensino Básico. A nossa análise incidiu sobre os aspectos que parecem explicar os resultados dos alunos, nos anos de 2003, de 2004 e de 2005, correspondendo, os dois primeiros anos, aos resultados das Provas de Aferição, e o último ano à primeira chamada dos exames de 9º ano. O estudo comparativo destes testes assentou na interpretação dos critérios de avaliação utilizados na cotação das respostas, no tipo de conteúdos testados e no tipo de itens apresentados aos alunos. Tanto esta interpretação como o estudo de caso de uma escola mostra que os indicadores estatísticos disponibilizados pelo *Gabinete de Avaliação Educacional* não permitem tirar ilações fiáveis e úteis sobre o desempenho dos alunos. As conclusões apontam graves erros na construção destes instrumentos de avaliação em termos da validade e da fiabilidade dos mesmos. Na verdade, os resultados dos alunos, nas várias provas, não são comparáveis e não oferecem bons indicadores do desempenho dos alunos.

Introdução

As Provas de Aferição (PA) têm sido realizadas em Portugal no final de cada ciclo do ensino básico, mais precisamente no final dos 4.º, 6.º e 9.º anos de escolaridade, nas disciplinas de Língua Portuguesa e de Matemática. As provas são aplicadas anualmente a uma amostra da população do respectivo ciclo escolhida criteriosamente de forma a ser representativa da população nacional de alunos e têm sido implementadas desde o ano 2000: «... o principal objectivo da realização das provas de aferição é o de fornecer informação sobre o desempenho dos alunos, face ao desenvolvimento de competências consideradas essenciais para cada ciclo de ensino. Neste sentido, permitem o levantamento de elementos de base para uma análise do funcionamento do sistema

1 Instituto Superior de Educação e Ciências

2 Instituto Superior de Educação e Ciências

3 Escola Secundária José Gomes Ferreira

educativo, podendo conduzir e alimentar o debate público sobre a Escola. Este debate deverá ser realizado pelos agentes educativos – professores, alunos, encarregados de educação, escola no seu geral e serviços centrais e regionais – como também por outros agentes sociais que partilham a preocupação comum de melhorar a eficácia do sistema educativo». (Ministério da Educação, 2004, p.7).

Em 2005, assistimos à realização de exames nacionais do 9º ano às mesmas disciplinas visadas nas PA; Matemática e Português. Embora não seja fácil descortinar no discurso político o objectivo dos exames, estes parecem servir para responsabilizar os alunos pelo seu desempenho. Enquanto que as PA foram implementadas com o objectivo de reunir informações sobre a eficácia do sistema educativo, nomeadamente sobre a eficácia do currículo nacional, os exames têm peso na avaliação global dos alunos (25% da nota final em 2005). No entanto, em ambas as formas de avaliação o importante é avaliar o resultado final das aprendizagens, o que corresponde a um tipo de avaliação sumativa pois ocorre no final de um ciclo de estudos (Valette, 1994).

Não obstante esta mudança na ponderação da avaliação sumativa, um relatório do Gabinete de Avaliação Educacional (GAVE) sobre os resultados do Exame de Matemática do 9º ano refere que tendo como referência «a percentagem da classificação média relativa à classificação máxima, por ano de aplicação», o desempenho dos alunos se tem mantido estável desde 2002 (Ministério da Educação, 2006). Dado que o desempenho dos alunos no exame corresponde a 71% de classificações negativas, de acordo com os dados apresentados neste relatório, os resultados das PA (aplicadas em 2002, 2003 e 2004 mas que não contaram para a avaliação final dos alunos) são idênticos aos resultados do exame de 2005. No entanto, uma leitura atenta dos resultados das PA não parece confirmar esta afirmação. Não foi ainda publicado um relatório idêntico referente aos resultados do Exame de Português, mas também nesta disciplina os resultados das PA não se afiguram idênticos aos do exame nacional. Tendo os alunos registado uma taxa de reprovação de apenas 23% no Exame de Português, este resultado é totalmente o oposto ao do Exame de Matemática.

A ausência de fiabilidade nas comparações entre provas, ou seja, a ideia de que um teste deve avaliar determinado desempenho de forma consistente ao longo do tempo, e que esse desempenho tem de ser testado da mesma forma em provas equivalentes (Creswell, 2003) motivou o presente estudo. Tendo como objectivo discutir as limitações das conclusões que foram avançadas pelo GAVE, através dos seus vários relatórios e veiculadas pelos meios de comunicação social, procurou-se descrever os indicadores que serviram de base a essas mesmas conclusões, acompanhando essa descrição de uma análise crítica. Apresenta-se ainda um estudo de caso relativo à participação de

uma escola neste tipo de análises onde é feita uma breve comparação entre os resultados de Matemática e de Língua Portuguesa.

As Provas de Aferição de Matemática

O tipo de avaliação que é feito nas PA assenta numa escala *qualitativa*, tendo em vista a recolha de dados relativos ao desempenho dos alunos por *competências* matemáticas e por *áreas temáticas*. Estudos que utilizam critérios qualitativos para a construção de escalas com níveis de desempenho diferentes também são feitos internacionalmente como, por exemplo, o estudo PISA (OECD, 2003). No caso português, a correcção das provas é planeada com antecedência juntando os professores correctores em vários momentos do processo de forma a assegurar a uniformização dos critérios de classificação. O conjunto de itens que forma o enunciado da prova é escolhido e é indexado a um conjunto de conteúdos e competências que se pretende avaliar. A cada item corresponde um critério de avaliação que permite ao corrector atribuir-lhe uma classificação qualitativa. Embora seja possível que os professores correctores avaliem de forma homogénea os resultados dos vários itens, atendendo ao Quadro 1 é notória a forma pouco fiável como são distribuídos os itens pelas competências e pelas áreas temáticas (Ministério da Educação, 2004, p. 151). Veja-se a pouca homogeneidade: Como se podem tirar, por exemplo, conclusões sobre a capacidade comunicativa dos alunos se o número de itens dedicado a esta competência é apenas de 3?

Quadro 1 - Distribuição dos itens na PA de Matemática do 9º ano de 2003

Predominância das áreas temáticas e tipos de competência					
Áreas temáticas	**Tipos de competência**				**Número de itens**
	Conhecer conceitos e procedimentos	Resolução de problemas	Raciocínio	Comunicação	
Números e Cálculo	4.1.1, 4.1.2, 10.1, 10.2, 12.1	4.2, 6.2, 12.2	4.3, 7, 9		11
Geometria	2.1, 2.2.1, 2.2.2, 5.2, 8	5.1	13		7
Estatística e Probabilidades	1.1, 1.2		6.3	6.1, 11	5
Álgebra e Funções	3.1, 3.2	3.3.2		3.3.1	4
Número de itens	**14**	**5**	**5**	**3**	**27**

Quadro 2 - Critério de avaliação de um item da PA de Matemática do 9º ano de 2003

11. Numa competição de natação sincronizada, cada exercício é avaliado por dois grupos de cinco juízes: um grupo avalia o Mérito Técnico e outro grupo a Impressão Artística. A nota final do exercício é calculada de acordo com as seguintes etapas:

1. Das cinco notas atribuídas por cada grupo de juízes, eliminam-se a nota mais baixa e a nota mais alta de cada grupo.
2. Calcula-se a média das restantes três notas atribuídas por cada grupo de juízes.
3. Utilizando as médias obtidas na etapa 2,
   - multiplica-se por 6 a média das notas atribuídas pelos juízes do Mérito Técnico;
   - multiplica-se por 4 a média das notas atribuídas pelos juízes da Impressão Artística.
4. A nota final do exercício é obtida pela soma dos valores obtidos na etapa 3.

Um concorrente obteve as seguintes notas num certo exercício:

Mérito Técnico	8,0	8,4	8,5	8,6	7,6
Impressão Artística	8,6	8,3	8,3	8,1	8,7

Calcula a nota final deste exercício, conforme as etapas descritas. Indica, em cada etapa, as decisões que tomares e apresenta os cálculos que efectuares.

Item 11

Na resolução do problema o aluno deverá seguir as seguintes etapas:

1. Eliminar as classificações extremas em cada categoria.	Mérito Técnico (MT): 8,6 e 7,6 Impressão Artística (IA): 8,1 e 8,7
2. Calcular a média dos restantes valores em cada categoria.	Média (MT): 8,3 Média (IA): 8,4
3. Multiplicar os valores obtidos pelo factor de ponderação.	8,3 × 6 = 49,8 8,4 × 4 = 33,6
4. Somar os valores obtidos e apresentar a resposta.	49,8 + 33,6 = 83,4

Resposta Correcta: 83,4.

3 Concretiza as quatro etapas de modo correcto, e há evidência de ter chegado à resposta correcta.

2 Concretiza correctamente apenas três das etapas [a].

2M Concretiza as três primeiras etapas apenas para uma das categorias.

1 Há algum trabalho, revelando alguma compreensão do problema [a].

1A Responde correctamente à pergunta, mas não apresenta o trabalho desenvolvido.

0 Apresenta outra resposta, além das mencionadas.

Nota:
(a) Deverá ser registada a letra **M** sempre que o aluno evidenciar conhecer o conceito de média aritmética.

Após a correcção das provas, os resultados são compilados sob a forma de um relatório anual e as avaliações dos alunos, quanto às suas competências e conhecimentos das diversas áreas temáticas, podem ser lidas graficamente (Ministério da Educação, 2004, p. 186).

A escala tem 4 graduações: Não Respondeu, Zero, Intermédio, Máximo. Isso significa que tanto o aluno que apresenta um tipo de resposta 90% satisfatória, como o aluno que apresenta um tipo de resposta 10% satisfatório, está no nível intermédio. Poderá argumentar-se que nas PA os conceitos de resposta 10% satisfatória ou de resposta 90% satisfatória não existem, no entanto, não deixa de transparecer uma escala curta e pouco detalhada. Ao contrário, o estudo Pisa, que também utiliza uma escala qualitativa, apresenta uma distribuição de itens muito mais homogénea, como se pode ver nos quadros 3 e 4. No estudo Pisa, os resultados baseiam-se num maior e mais homogéneo número de itens, que são construídos com base numa escala com indicadores muito mais descritivos, que permitem discriminar os diferentes níveis de desempenho (OECD, 2003, pp. 334 e 47).

Quadro 3 - Distribuição dos itens no Pisa 2003

Distribution of items by the dimensions of the PISA framework for the assessment of mathematics

	Number of items[1]	Number of multiple-choice items	Number of complex multiple-choice items	Number of closed-constructed response items	Number of open-constructed response items	Number of short response items
Distribution of mathematics items by "overarching ideas"						
Space and shape	20	4	4	6	4	2
Change and relationships	22	1	2	4	11	4
Quantity	23	4	2	2	1	14
Uncertainty	20	8	3	1	5	3
Total	*85*	*17*	*11*	*13*	*21*	*23*
Distribution of mathematics items by competency cluster						
Reproduction	26	7	0	7	3	9
Connections	40	5	9	4	9	13
Reflection	19	5	2	2	9	1
Total	*85*	*17*	*11*	*13*	*21*	*23*
Distribution of mathematics items by situations or contexts						
Personal	18	5	3	1	3	6
Educational/Occupational	20	2	4	6	2	6
Public	29	8	2	4	8	7
Scientific	18	2	2	2	8	4
Total	*85*	*17*	*11*	*13*	*21*	*23*

O critério de fiabilidade, que anteriormente mencionámos, está patente no estudo PISA, porque foi seguido um outro critério-chave que deve guiar a elaboração de instrumentos de avaliação - a validade (Haladyna, 1994). Esta só é assegurada quando os itens reflectem aquilo que se pretende avaliar. Por outras palavras, são os itens representativos dos conteúdos ou aprendizagens que se pretendem avaliar? São elaborados itens suficientes para medir o desempenho dos alunos por área temática e por competência?

Como acabámos de mostrar através da comparação entre as escalas de classificação e a distribuição de itens utilizadas nas PA e no estudo PISA, a validade, e portanto, a fiabilidade, estão comprometidas nas PA. Em particular, as PA não incluem itens suficientes por área temática, e por competência, de modo a permitir conclusões fiáveis sobre o desempenho dos alunos.

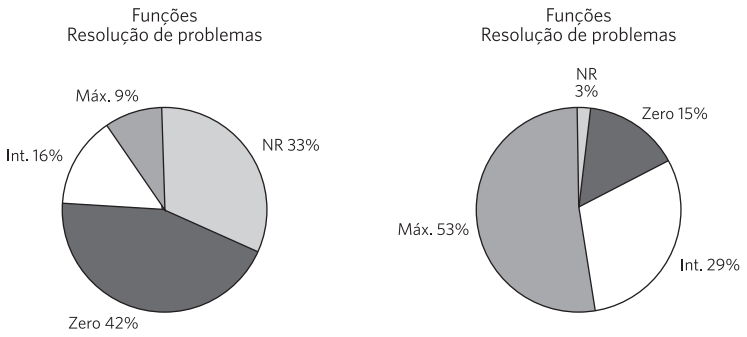
Quadro 4 - Escala usada para a classificação dos itens do Pisa 2003

Level	WHAT STUDENTS CAN TYPICALLY DO
6	At Level 6, students can conceptualise, generalise, and utilise information based on their investigations and modelling of complex problem situations. They can link different information sources and representations and flexibly translate among them. Students at this level are capable of advanced mathematical thinking and reasoning. These students can apply this insight and understanding, along with a mastery of symbolic and formal mathematical operations and relationships, to develop new approaches and strategies for attacking novel situations. Students at this level can formulate and precisely communicate their actions and reflections regarding their findings, interpretations, arguments, and the appropriateness of these to the original situations.
5	At Level 5, students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare, and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well-developed thinking and reasoning skills, appropriately linked representations, symbolic and formal characterisations, and insight pertaining to these situations. They can reflect on their actions and can formulate and communicate their interpretations and reasoning.
4	At Level 4, students can work effectively with explicit models for complex concrete situations that may involve constraints or call for making assumptions. They can select and integrate different representations, including symbolic ones, linking them directly to aspects of real-world situations. Students at this level can utilise well-developed skills and reason flexibly, with some insight, in these contexts. They can construct and communicate explanations and arguments based on their interpretations, arguments and actions.
3	At Level 3, students can execute clearly described procedures, including those that require sequential decisions. They can select and apply simple problem-solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They can develop short communications reporting their interpretations, results and reasoning.
2	At Level 2, students can interpret and recognise situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures or conventions. They are capable of direct reasoning and making literal interpretations of the results.
1	At Level 1, students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are obvious and follow immediately from the given stimuli.

A utilização das PA para comparar anos diferentes

Embora os resultados das PA sejam mencionados pelos *media* para comparar desempenhos de anos diferentes, o que se constata é que este tipo de comparação não é viável pois incorre em erros graves. Repare-se na figura seguinte, que diz respeito a um exemplo de comparação de resultados da competência "resolução de problemas" associada à área "funções" , nos anos de 2003 (Ministério da Educação, 2004, p. 157) e de 2004 (Ministério da Educação, 2006, p. 151).

Quadro 5 - Funções/Resolução de Problemas (2003 à esquerda e 2004 à direita)

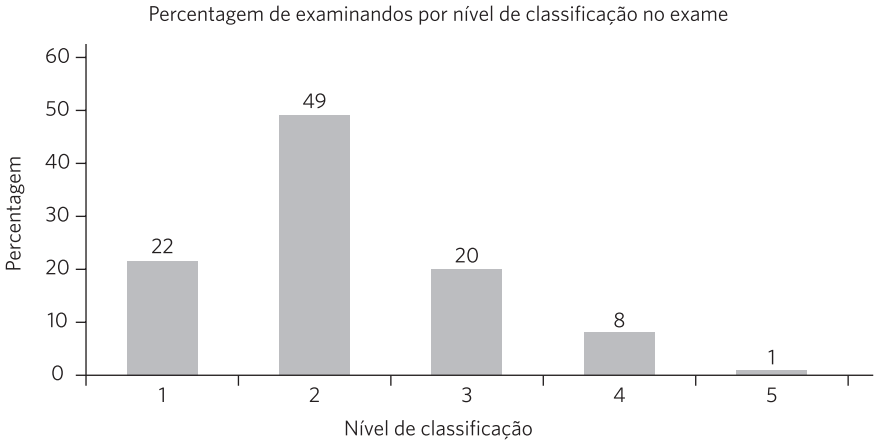


É manifesto que a comparação apresenta diferenças difíceis de justificar, dado que não houve, de um ano para o outro, nenhuma alteração radical no plano curricular ou no âmbito das políticas educativas vigentes. No entanto, podemos constatar nos relatórios das PA inúmeros incidentes semelhantes a este, bastando para tal cruzar os dados relativos a outras áreas temáticas/competências (Ministério da Educação, 2004; Ministério da Educação, 2006). As causas podem ser várias: poucas questões provocam pouca fiabilidade nos resultados, a diferença de graus de dificuldade das perguntas apresenta diferenças de um ano para o outro e a escala utilizada é pouco discriminativa. O que é certo é que não podemos concluir que houve, de um ano para o outro, uma melhoria na resolução de problemas associada a questões de funções. O que podemos concluir é que as PA de 2003 e de 2004 não são equivalentes e, como tal, os resultados dos alunos não são comparáveis nos dois anos. Sendo assim, as provas podem ter outras finalidades como, por exemplo, comparar desempenhos em diferentes zonas do país, no mesmo ano, mas certamente que não pode ter a de comparar desempenhos de anos diferentes.

Comparação das PA com os Exames Nacionais do 9ºano

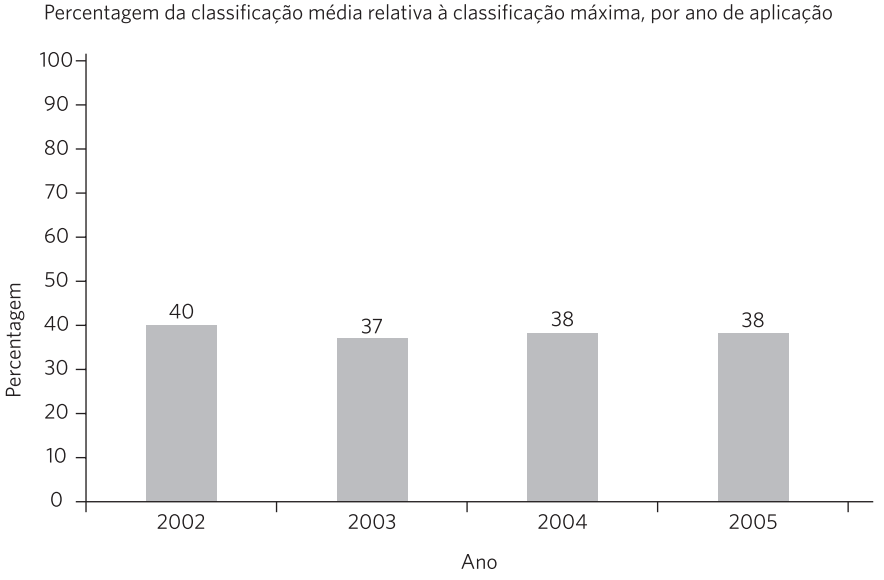
Quando se efectuaram os Exames Nacionais do 9º ano, em 2005, observou-se que os resultados de Matemática foram profundamente desanimadores - 71% de classificações negativas (GAVE, 2006, p.8).

Quadro 6 - Resultados do Exame de Matemática do 9º Ano de 2005



Face a resultados tão maus, imediatamente se levantaram inúmeras opiniões. Muitos intervenientes nestas discussões, incluindo o GAVE, recorreram a PA de anos anteriores com o objectivo de fazer uma comparação. Contudo, este tipo de comparação levanta várias questões: será legítimo comparar estas provas de índole diferente? Em caso afirmativo, que indicador estatístico se deve usar?

Quadro 7 - Comparação das Provas de Aferição/Exame Nacional



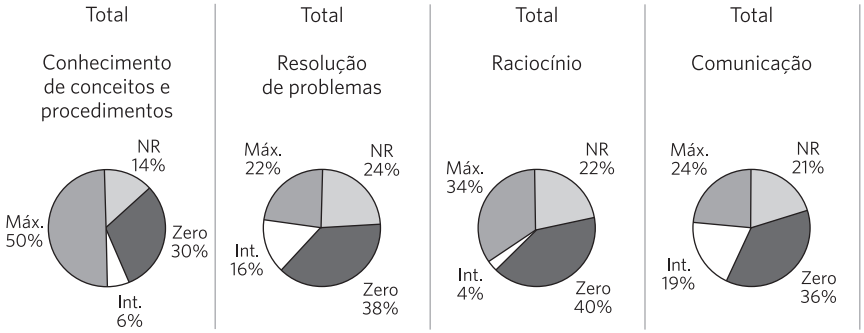
A primeira constatação que pode fazer-se é que o valor do indicador global de desempenho se tem mantido relativamente estável, embora preocupantemente baixo.

Atendendo a que uma PA é de avaliação qualitativa e um exame é de avaliação quantitativa esse facto levanta imediatamente um problema de escala. O relatório oficial com os resultados do Exame do 9º ano de Matemática de 2005 apresenta o seguinte gráfico relativo a uma comparação dos resultados das PA de anos anteriores com os do Exame (GAVE, 2006, p.17).

O problema de encontrar um indicador estatístico que permita uma comparação não pode ser facilmente resolvido ao comparar provas de índole qualitativa com provas de índole quantitativa. O indicador utilizado foi a percentagem da classificação média relativa à classificação máxima, por ano de aplicação. Está assim a considerar-se para as PA uma "classificação média" expressa em termos numéricos, quando se sabe que os números expressos nas classificações das perguntas representam graus qualitativos e não uma grandeza numérica. Por isso, este indicador não parece cumprir o seu propósito. Apesar das deficiências expressas, as diversas comparações foram divulgadas pelos *media*. Por exemplo, no jornal *Público* de 17 de Janeiro de 2006, apareceu a seguinte notícia (p.23): «... os desempenhos dos alunos que em Junho de 2005 realizaram, pela primeira vez, o exame nacional de Matemática do 9º ano, foram "muito fracos", aliás na continuidade dos relativos às provas de Aferição do 3º ciclo (que testavam os mesmos conhecimentos mas não contavam para nota). O relatório do Gabinete de Avaliação Educacional (GAVE), organismo responsável pela elaboração dos testes, está concluído e vem confirmar muitas das dificuldades já diagnosticadas. De uma forma geral, escreve o GAVE, o desempenho médio dos alunos "nunca se apresentou nem bom nem muito bom" em qualquer das competências ou domínios temáticos analisados. Mas onde os estudantes portugueses revelam maiores dificuldades é na competência relativa à resolução de problemas. "Mesmo os problemas mais simples" ».

Estas conclusões quanto às competências levantam ainda outras dúvidas. A própria forma como é lida a informação por parte dos agentes que emitiram estas conclusões não é facilmente compreendida quando se lê os relatórios. Não parece credível, atendendo ao desempenho dos alunos nas várias competências matemáticas, inferir que foi na resolução de problemas que os alunos apresentaram piores resultados. Observe-se os resultados totais da PA de 2004, quanto às competências (Ministério da Educação, 2006, p.150):

Quadro 8 - Totais da prova de aferição de 2004, por competências



É certo que a resolução de problemas apresenta maus resultados, mas acontece o mesmo com o raciocínio e a comunicação. De facto, a informação disponível e respectivas interpretações são falaciosas e em nada contribuem para fornecer informação sobre o desempenho dos alunos.

O caso de uma Escola Secundária com Ensino Básico (ESEB)

Em Dezembro de 2005, o GAVE solicitou a todas as escolas com Ensino Básico que fosse realizada uma análise dos exames nacionais de Matemática do 9º Ano (1ª chamada), do ano lectivo de 2004/2005, tendo em conta os resultados globais nacionais obtidos neste exame e os resultados da escola. Este estudo pretendia, entre outras coisas, comparar os resultados nacionais com os obtidos na escola, tanto a nível global como por área temática/competência.

Para o efeito, foram fornecidos às escolas, por parte do GAVE, os resultados nacionais e os resultados por escola, tanto a nível global como por áreas temáticas (Números e Cálculo, Estatística e Probabilidades, Álgebra e Funções e Geometria), e por competências matemáticas (Conceitos e Procedimentos, Raciocínio, Resolução de Problemas, Comunicação). No entanto, o GAVE não forneceu quaisquer elementos sobre o significado das percentagens nem sobre a escala utilizada. Sendo assim, a ESEB ficou sem saber como foram atribuídas as classificações *máximo*, *intermédio*, *não respondeu* e *zero* às respostas dos alunos. Assim, para poder efectuar a análise solicitada, a ESEB pediu esclarecimentos ao GAVE quanto à forma como foram atribuídas as classificações.

A partir dos esclarecimentos fornecidos pelo GAVE ficou a saber-se que as percentagens relativas às classificações têm por base o total de itens de todas as provas realiza-

das. A 1ª chamada do Exame de Matemática do 9º ano de 2005 foi realizada por 84788 alunos. Sendo assim, o total de itens pode ser obtido multiplicando o número de alunos que efectuaram a prova pelo número de itens da prova. Uma vez que o número de itens da prova é igual a 18, o total de itens nacional é igual a 84.788 x 18=1.526.144. Quando, por exemplo, se vê que 34% dos alunos obtiveram resposta máxima, esta percentagem recai sobre 1.526.144 itens. Sabendo que esta era a regra, a ESEB calculou, para o total de itens interno, o número 80 x 18=1440, correspondente a 80 provas.

Com base nos valores incluídos nos gráficos circulares representados no Quadro 10, correspondentes aos resultados globais nacionais e aos resultados da ESEB, o GAVE fez a seguinte solicitação: "A partir dos resultados globais, registe duas possíveis causas que possam explicar os resultados obtidos pela escola."

Quadro 9 - Resultados globais obtidos na 1ª chamada do Exame de Matemática do 9º ano (2004/2005)

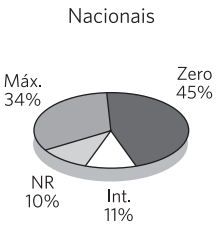


Gráfico 1

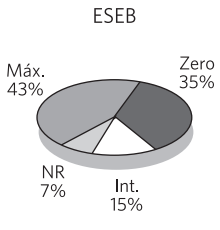


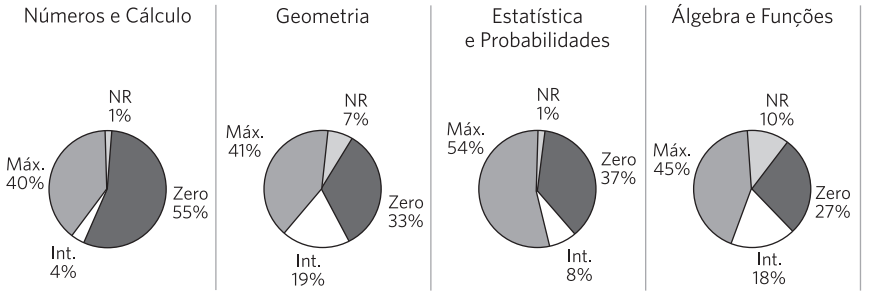
Gráfico 2

A resposta da ESEB salientou os seguintes aspectos:

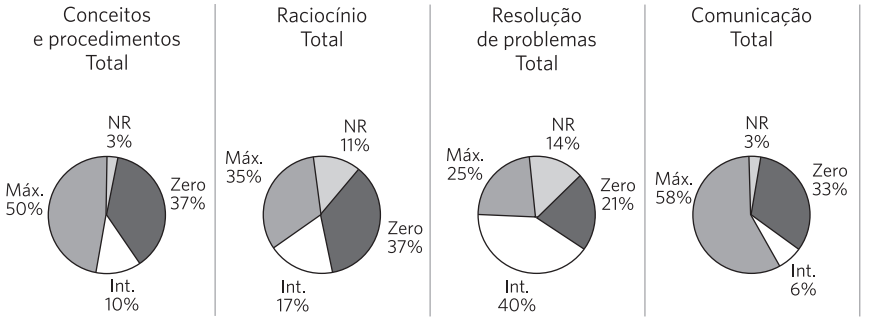
"O facto de se incluírem todas as respostas entre zero e o nível máximo, no nível intermédio, não permite estabelecer critérios de validade e de fiabilidade a este estudo, por neste nível estarem abrangidas respostas *quase certas* e respostas *praticamente erradas*. A circunstância de não estarem claros os aspectos relativamente aos quais se pretende estabelecer a comparação torna difícil a realização da análise proposta. Não estão reunidas as condições para a identificação de indicadores que possam contribuir para os objectivos do estudo proposto pelo GAVE."

Além da análise global, o GAVE solicitou ainda informação quanto aos resultados por áreas temáticas/competências, com base nos quadros 11, 12, 13 e 14: "Registe as áreas em que os alunos tiveram melhor desempenho e pior desempenho. Registe os aspectos das competências matemáticas em que os alunos tiveram melhor desempenho e pior desempenho. Registe duas causas que possam explicar os piores desempenhos."

Quadro 10 - Resultados da escola por Áreas temáticas.



Quadro 11 - Resultados da escola por Competências matemáticas.



Quadro 12 – Distribuição de itens da 1ª chamada do Exame de Matemática do 9º ano (2004/2005).

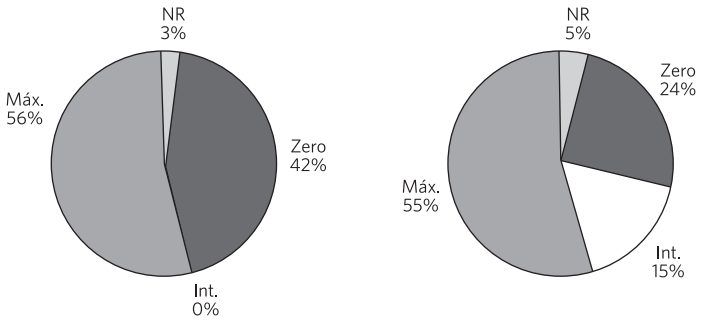
Competências / Áreas de Conteúdo	Conceitos e Procedimentos	Raciocínio	Resolução de problemas	Comunicação	Total de códigos
Números e Cálculo	2.1., 6., 10.				15
Estatística e Probabilidades	1.	4.			10
Álgebra e Funções	2.2.	8.2., 11.	5.2.	3.1., 3.2.	36
Geometria	5.1.1., 5.1.2., 7.1., 8.1.	7.2., 7.3.	9.		39
Total de códigos	44	32	16	8	

Quadro 13 – Resultados da ESEB – áreas temáticas/competências.

Competências / Áreas de Conteúdo	Conceitos e Procedimentos	Raciocínio	Resolução de problemas	Comunicação	Total
Números e Cálculo	41.8%				41.8%
Estatística e Probabilidades	45.0%	75.2%			63.1%
Álgebra e Funções	57.5%	41.0%	50.8%	61.3%	50.9%
Geometria	67.8%	23.5%	48.9%		49.2%
Total	55.3%	40.3%	49.8%	61.3%	

Ao preparar a resposta, a ESEB procurou, em primeiro lugar, compreender o que se entende por pior ou melhor desempenho. Três hipóteses se levantaram: a) a área/competência com melhor ou pior desempenho é a área/competência com maior ou pior percentagem de itens com cotação máxima (Quadros 10 e 11). Para se entender melhor os problemas desta definição, observe-se o seguinte:

Quadro 14 – Exemplo de comparação de desempenhos.



De acordo com os dados apresentados, independentemente da área ou competência que se pretenda comparar, segundo esta definição, o gráfico da esquerda apresentaria um melhor desempenho, por ter maior percentagem de respostas com cotação máxima. No entanto, dada a significativa percentagem de respostas intermédias no gráfico da direita, mais facilmente se atribuiria um melhor desempenho ao da direita.

b) a área/competência com melhor ou pior desempenho é a área/competência com maior ou menor número de itens com cotação máxima em *valores absolutos* (Quadros

10 e 11). Esta definição apresenta um problema de validade. Por exemplo, atendendo ao Quadro 12, o número total de respostas com cotação máxima em *Álgebra e Funções* recai sobre 6 itens, enquanto para *Estatística e Probabilidades* recai só sobre 2.

c) a área/competência com melhor ou pior desempenho é a área/competência que obteve melhores ou piores resultados em termos percentuais (Quadros 12 e 13). Por exemplo, quanto a esta última definição, quando se lê que os alunos da escola obtiveram 41% no cruzamento *Álgebra* e *Funções/Raciocínio*, este valor recai apenas sobre dois itens. Para se obter esta percentagem, foi necessário ver, com auxílio do quadro 13, quantos valores correspondem (classificação máxima) aos dois itens do cruzamento *Álgebra* e *Funções/Raciocínio*. Também foi necessário calcular a classificação média obtida pelos alunos nas questões em causa, e em seguida calcular a percentagem da classificação média em relação à classificação máxima. Permanecem subjacentes os mesmos problemas de validade dada a distribuição variável do número de itens por área/competência.

A resposta da ESEB focou os seguintes aspectos: " Em relação à avaliação do desempenho dos alunos por áreas temáticas, se o critério escolhido for a percentagem de número de *itens* com cotação máxima, verifica-se que é o pior desempenho que vai ter o maior número de *itens* com resposta máxima em valores absolutos; se o critério tiver por base o Quadro 13, verifica-se que não existe diferença significativa entre os considerados "melhor" e "pior" desempenhos. No que se refere à avaliação do desempenho dos alunos por competências matemáticas, verifica-se e reforça-se a tendência anteriormente referida para as áreas temáticas. Assim o critério de melhor/pior desempenho não parece estar bem definido, não sendo possível encontrar causas para explicar os piores resultados." A nosso ver, os indicadores apresentam deficiências injustificáveis do ponto de vista da fiabilidade da avaliação educacional em Portugal, constituindo este estudo de caso um exemplo da incerteza vivida pelas várias escolas que, a nível nacional, tiveram de proceder a esta análise.

0 caso do Português

Os resultados dos testes de Português, nos três anos em que se centra esta análise, e ainda no ano de 2006, apresentam um grau de variabilidade superior ao verificado para a Matemática, apesar dos resultados globais serem melhores a Português. Em 2005, o Português aparece em situação oposta à da Matemática, com apenas 23% dos alunos a obterem uma classificação negativa e 77% a obterem classificação positiva (Diário Digital, 2005). A situação para o Português é tanto mais gravosa que, ao contrário da

pequena oscilação de 6 pontos percentuais em Matemática, entre o exame de 2005 e de 2006, se verificou uma oscilação de 23 pontos percentuais nos exames de Português nestes dois anos. Enquanto que em 2005, 77% dos alunos obteve classificação positiva, em 2006 essa percentagem desceu para 54% (Diário Digital, 2006). Responsáveis pelo Ministério da Educação afirmaram que, apesar de o Português ter piorado um pouco, os resultados estavam dentro do intervalo esperado (Diário Digital, 2006).

No entanto, tal discrepância põe em causa o princípio de fiabilidade, sendo que seria de esperar que uma população análoga tivesse obtido resultados semelhantes em provas equivalentes com conteúdos semelhantes. Verifica-se igual discrepância na comparação entre os resultados do Exame de Português de 2005 e das PA de 2003 e de 2004 a esta disciplina, assim como se afigura uma anomalia que os primeiros exames de 2005 revelem uma diferença tão acentuada entre os resultados dos alunos a Português e a Matemática.

Passamos assim, para efeitos de comparação entre os vários testes realizados a Matemática e a Português, a mostrar como os resultados das PA e do Exame de Português de 2005 não são idênticos de ano para ano, e como os resultados não deveriam ter sido tão díspares entre Matemática e Português nos primeiros exames realizados em 2005. Primeiro, as PA a Português não são idênticas entre si e não são idênticas ao Exame Nacional. Quando se considera, por exemplo, o tipo de competências testadas na área da *Compreensão da Leitura*, nos anos de 2003 e de 2004, verifica-se que as mesmas não se mantiveram constantes de ano para ano (Ministério da Educação, 2004, p. 131; Ministério da Educação, 2006, p.126).

Quadro 15 – Competências na Área da Compreensão da Leitura, 2003

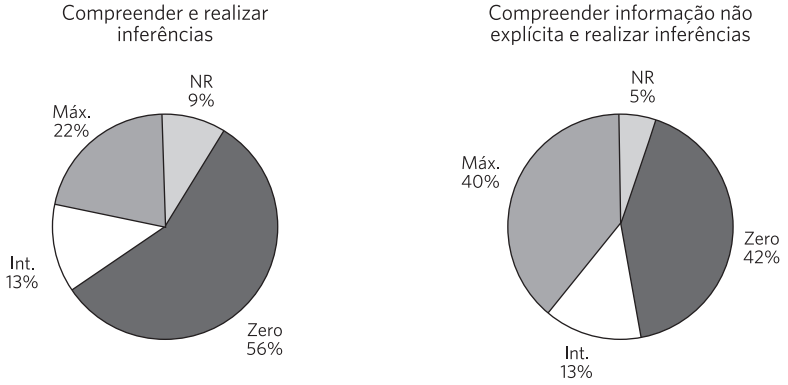
Operacionalização da competência	Itens
Reconhecer e identificar a informação solicitada com as palavras do texto.	1, 6, 7
Compreender a informação explícita no texto e responder por palavras diferentes (realizar paráfrases).	2, 5, 11
Compreender a informação não explícita no texto (realizar inferências).	3, 4, 10, 12
Reconhecer as marcas do narrador e transformar o discurso.	9
Expressar opiniões pessoais fundamentadas	8

Quadro 16 - Competências na Área da Compreensão da Leitura, 2004

Operacionalização da Competência	Itens
Reconhecer e identificar a informação solicitada e responder com as palavras do texto	10
Compreender a informação explícita no texto e responder por palavras diferentes	2, 5, 10
Compreender a informação não explícita no texto (realizar inferências)	1, 3, 4, 6, 7, 9
Conhecer esquemas rimáticos	8
Expressar opiniões fundamentadas	7, 9
Extrair informação a partir de enunciados não verbais	2, 10

Como os quadros mostram, a competência "Reconhecer as marcas do narrador e transformar o discurso" é testada em 2003, mas desaparece em 2004, para dar lugar a duas competências diferentes, a saber; "Conhecer esquemas rimáticos" e "Extrair informação a partir de enunciados não verbais," o que introduz um factor de variabilidade adicional.

Quadro 17 – Compreensão da leitura: à esquerda PA de 2003 e à direita PA de 2004.

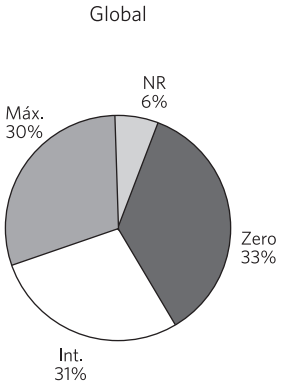


É de salientar que, tal como em Matemática, a distribuição de itens pelas várias competências no caso do Português também é muito variável de ano para ano, o que compromete qualquer interpretação comparativa quanto ao pior ou melhor desempenho dos alunos de um ano para o outro. Como sugerem os relatórios das PA, uma leitura global dos resultados levar-nos-ia a considerar que os resultados se mantiveram constantes, mas uma análise mais atenta revela que dentro das três áreas testadas há grande variabilidade nos resultados. Por exemplo, a competência "Compreender e realizar inferências", que em 2003 foi testada com quatro itens, e que em 2004 aparece como "Compreender informação não explícita e realizar inferências", que foi testada com seis

itens, regista uma enorme diferença percentual (Ministério da Educação, 2004, p. 134; Ministério da Educação, 2006, p. 129).

Como os gráficos mostram, houve 56% de respostas zero em 2003 contra apenas 42% em 2004. Verificam-se diferenças semelhantes nas outras competências dentro da área da *Compreensão da Leitura*, bem como em competências testadas nas áreas do *Conhecimento Explícito da Língua* e da *Composição Escrita Compositiva*", apresentando assim as provas de Português os mesmos erros de construção que as provas de Matemática. Os resultados no Exame de Português de 2005 também não parecem congruentes com os resultados das PA. Note-se que o somatório dos níveis médio e intermédio é de 61% na PA de 2003, e que no exame de 2005 os alunos atingiram uma taxa de aprovação de 77%.

Quadro 18 – Total Global Língua Portuguesa/PA 2003



Como temos vindo a discutir, é problemático converter um indicador qualitativo num indicador quantitativo para estabelecer comparações entre as PA e os exames. Mas é difícil de explicar um resultado tão bom no Exame de Português, face aos dados relativos às PA realizadas em anos anteriores. Na verdade, este resultado não parece dever-se unicamente às diferentes escalas utilizadas. O próprio conteúdo do exame foi muito diferente do que foi testado nas PA em 2003 e 2004. Os melhores resultados dos alunos no exame podem dever-se, em parte, à predominância de perguntas que requerem um nível de compreensão elementar, e de perguntas que requerem apenas um conhecimento implícito sobre o funcionamento da língua.

É igualmente difícil de explicar a diferença entre os resultados dos exames de Português e de Matemática. Se considerarmos os resultados globais das PA a Matemática e a

Português, e fizermos o somatório dos níveis de desempenho intermédio e máximo, as provas de 2003 e de 2004 apontam para uma diferença de 15 pontos percentuais entre a Matemática e o Português (*cf.* Ministério da Educação, 2004, p. 185).

De acordo com estes dados, seria de esperar uma diferença semelhante nos exames de 2005 e não uma taxa de reprovação de 71% a Matemática, e uma taxa de reprovação de 23% a Português. Acresce que, segundo dados da Inspecção-Geral de Educação, a percentagem de alunos do 9º ano, que em anos anteriores reprovaram a Matemática e a Português na avaliação interna das escolas, foi de cerca de 33% e de 20%, respectivamente; uma diferença entre disciplinas muito mais próxima daquela que consta dos resultados das PA do que dos resultados dos exames de 2005 (Portal do Governo, 2005).

Conclusão

Os resultados das Provas de Aferição (PA) não são comparáveis entre si nem são comparáveis aos exames realizados em 2005. A ausência de validade dos testes compromete a fiabilidade; ou seja, não podemos estar certos que uma população análoga de alunos obtenha resultados semelhantes pois as provas não são equivalentes. No caso das PA, podemos concluir que a forma como são feitas não permite atingir os três objectivos preconizados aquando da sua implementação, a saber; analisar o funcionamento do sistema educativo, conduzir o debate público e melhorar a eficácia do sistema educativo.

Estes objectivos só poderiam ser atingidos se as provas dessem indicadores fiáveis sobre as áreas e as competências em que os alunos têm pior e melhor desempenho. Ao invés, a comunicação social tem lançado, com base nos documentos do GAVE e nas declarações de dirigentes políticos, um debate público que distorce a realidade e que em nada contribui para a melhoria do sistema educativo. Na ausência de indicadores fiáveis é impossível fazer uma leitura longitudinal dos resultados e, com base neles, melhorar o sistema educativo.

Parece-nos ainda que os exames, por não se revelarem consistentes com os resultados de anteriores Provas de Aferição, para além de comprometerem a avaliação conduzida pelo Gabinete de Avaliação Educacional (GAVE), podem dar uma mensagem errada aos alunos. Afinal, parece que os resultados nos exames são como jogar à sorte; são melhores ou piores conforme o ano. Que as coisas vão mal com as aprendizagens dos alunos a Matemática e a Português já os estudos internacionais PISA (OCDE, 2000 e

2003) tinham revelado, não acusando qualquer diferença significativa, quer a Matemática, quer a Língua Materna, de triénio para triénio. Uma vez que as PA foram substituídas pelos exames, o PISA, no qual Portugal voltou a participar em 2006, pode servir de referência longitudinal para avaliar o sistema educativo. No entanto, permanece a questão da mensagem que queremos dar aos alunos. Uma mensagem honesta, que transmita a ideia de que se devem preparar e estudar os conteúdos que foram trabalhados ao longo do ano para obterem bons resultados, só pode passar se os exames forem bem construídos.

Bibliografia

Creswell, J. (2003). *Research design: Qualitative, quantitative and mixed approaches*. London: Sage Publications.

Haladyna, T. (1994). *Developing and validating multiple-choice test items*. New Jersey: Lawrence Erlbaum Publishers.

Ministério da Educação (2004). *Provas de aferição do ensino básico 4.º 6.º 9.º anos – 2003*.

Ministério da Educação (2006). *Provas de aferição do ensino básico 4.º 6.º 9.º anos – 2004*.

GAVE/Ministério da Educação (2006). *Resultados do exame de matemática do 9ºano 2005 1ª chamada*. Consultado em Janeiro de 2007, http://www.gave.pt/2005/basico/relatorio_9ano_matematica_2005 (pdf).

OECD (Organisation for Economic Co-operation and Development) (2003). Education at a Glance. Paris: OECD.

OECD (Organisation for Economic Co-operation and Development) (2000). *Education at a Glance*. Paris: OECD.

Portal do Governo (2005). Divulgação pública dos resultados do PISA-2003 e de medidas para melhorar o ensino da matemática. Consultado em Março de 2005, http://www.portugal.gov.pt/Portal/PT/Governos/Governos_Constitucionais/GC17.

Público (2006). Dificuldades na resolução de problemas. Jornal *Público*, 17 Janeiro, p.23.

Valette, R. (1994). Teaching, testing and assessment. *In* C. Hancock (Ed.), *Teaching, testing and assessment* (pp.1-39). Chicago: National Textbook Company.

Diário Digital/Lusa (2005). Consultado em Janeiro de 2007, http://diariodigital.sapo.pt/news.asp?section_id=61&id_news=182846.

Diário Digital/Lusa (2006). Consultado em Janeiro de 2007, http://diariodigital.sapo.pt/news.asp?section_id=61&id_news=235913.

Résumé

Cet article présente une analyse critique des résultats des tests de compétences et des examens de mathématiques et de portugais de la neuvième année (en France correspont à la quatrième année)de l`Enseignement Basique.Notre analyse a mis en évidence les aspects qui semblent expliquer les résultats obtenus par les élèves des années 2003, 2004 et 2005, correspondant les deux

premières années aux résultats des tests de compétences et le dernier au premier appel des examens de la neuvième (quatrième année).L`étude comparatif de ces tests s`est fondée sur l`interprétation des critères d`évaluation utilisés dans la cotation des réponses, sur le type de contenus testés et sur le type d`itens présentés aux élèves. Soit cette interprétation soit l`étude de cas d`une école confirment que les indicateurs statistiques disponibilisés par le Bureau d`Évaluation Éducationnelle ne permettent pas d`obtenir des conclusions fiables et utiles sur le travail des élèves.

On a conclu qu`il y a eu de graves erreurs dans l` élaboration de ces instruments en ce qui concerne la validité et la fiabilité des mêmes.

En effet, les résultats obtenus dans les différentes épreuves ne sont pas comparables et donc ne sont pas de bons indicateurs du travail des élèves.

Abstract

This article presents a critical analysis of the results of the Portuguese assessment tests and of the Portuguese national math and Portuguese exams at the 9th grade level. Our analysis focuses on the variables that seem to explain the results students obtained in the years 2003, 2004 and 2005, with the first two years corresponding to the assessment tests and the last year corresponding to the exam results. The comparison between these two types of tests includes the interpretation of the evaluation criteria utilized to score the students' responses, of the type of content tested and of the types of items presented to the students. This comparison exercise, as well as the case study of a school, shows that the statistical indicators gathered by the Educational Evaluation Services of the Ministry of Education do not allow reliable inferences about students´ achievement. The conclusions point out several errors in the development of these assessment tools in terms of their validity and reliability. In short, the students' results in the tests discussed are not comparable and do not offer good indicators of students' performance.