4
Recorreção, análise dos testes e resultados

No âmbito das avaliações em larga escala, a introdução da escrita na fase da alfabetização é algo extremamente recente no Brasil, tendo sido implementada pelo CEALE-UFMG.

As primeiras experiências nesse campo ocorreram no Proalfa, nos anos de 2006 e 2007, e também no pré-teste da Provinha Brasil, em 2007. Contudo, tais experiências não resultaram, naquele momento, na geração de uma escala de proficiência em escrita, devido ao pouco conhecimento acerca do tratamento estatístico de informações oriundas desse tipo de avaliação.

Nesse sentido, o PAEBES-Alfa pode ser considerado um avanço no âmbito das avaliações em larga escala voltadas para a alfabetização. Além de avaliar, a partir de 2009, todos os alunos matriculados no 1º ano/1ª série até o 3º ano do Ensino Fundamental das redes públicas do Estado do Espírito em leitura, compreensão e escrita, os resultados dessas avaliações dão origem a uma escala de desempenho da leitura e da escrita.

Na construção da escala de proficiência em escrita e sua interpretação, como visto nos capítulos anteriores, para cada um dos itens foram elaboradas chaves de correção que orientaram o processo de avaliação da produção escrita dos alunos participantes do PAEBES-Alfa.

Como buscamos analisar os instrumentos e procedimentos utilizados na avaliação da escrita, neste capítulo apresentamos tanto os diferentes procedimentos utilizados na correção dos itens de escrita, como os procedimentos e resultados da recorreção, e ainda uma análise dos testes, considerando a dificuldade dos itens.

4.1
A correção no PAEBES-Alfa

A avaliação da produção escrita do PAEBES-Alfa, de 2009 a 2011, era realizada, presencialmente, por um conjunto bastante grande de corretores,

divididos em três turnos de trabalho de correção, cada turno de com cinco horas de duração. Esses corretores, aproximadamente 400, organizados em células (mesas de trabalho), não só avaliavam a produção escrita, com base na chave de correção, como também transcreviam para o cartão de resposta a marcação feita pelos alunos nos itens de múltipla escolha.

Em cada turno de trabalho, as células eram monitoradas por um supervisor de célula. Ele era responsável pela garantia da distribuição de tarefas para todos os corretores, pela manutenção da produtividade e por coletar dúvidas referentes ao processo de correção (habilidade avaliada, legibilidade da letra do aluno, adequação da gradação da chave à resposta do aluno), levando-as ao supervisor de turno que respondia por todo o processo, inclusive o pedagógico.

Nesse processo de correção, apesar de um corretor ter sido o responsável pela avaliação de todos os testes de uma mesma turma, não havia dado disponível que permitisse associar esse corretor à turma ou ao aluno cujo teste tinha corrigido. A digitalização das imagens dos testes dos alunos foi realizada posteriormente, para fins de estudo e do projeto de constituição de corpus de escrita infantil, coordenado pela professora Gladys Rocha, do CEALE-UFMG.

Em 2012, o processo de correção passou a ser via sistema *on-line*, o que exigiu a digitalização prévia dos testes, a fim de serem disponibilizados para as equipes de correção.

Nesse ano, foram capacitados para a avaliação 82 corretores, divididos em duas supervisões. Assim, de modo semelhante ao que era realizado na correção presencial, os corretores foram divididos em supervisões, cuja finalidade era monitorar: (i) o andamento dos trabalhos para garantir o cumprimento de prazos; (ii) a coerência da correção realizada por cada um dos corretores alocados sob sua responsabilidade e (iii) resolver dúvidas relativas aos aspectos pedagógicos e resolver possíveis discrepâncias.

Essa modalidade de correção se constitui em um avanço, pois como cada questão traz um sequencial que permite associá-la ao aluno e cada corretor tem um código de identificação, ao designar os itens para um dado corretor, é possível, posteriormente, recuperar essa informação e associar corretor ao item/aluno.

4.2
A recorreção do grupo Estudo

A proposição de uma nova correção dos testes do grupo de alunos selecionado partiu de uma das hipóteses desta tese, segundo a qual a queda da proficiência desses alunos poderia ser decorrente da interferência da subjetividade do corretor.

Uma vez que para a produção dos resultados são utilizados modelagens da TRI, que levam em consideração apenas dois componentes ou facetas: respondentes e itens do teste, em uma situação de avaliação de produção escrita coloca-se em jogo outra faceta que pode interferir no resultado alçando pelo aluno: o corretor.

Conforme aponta Eckes (2011), o modo como o corretor se comporta diante da correção, ou seja, se ele é severo, leniente ou ainda se provoca um efeito alo[1], pode interferir no resultado final dos textos produzidos pelos alunos.

Mas como, ainda hoje, no Brasil, não se trabalha com o modelagem de Rasch com multifacetas (*Many-Facet Rasch Measurement* –MRFM), que permite, por exemplo, avaliações multidimensionais, interpretações de resultados educacionais com utilização de subescalas e estimativa da influência dos corretores nas pontuações dos respondentes em itens de resposta graduada (Eckes, 2011), buscamos, assim, com uma nova correção dos itens do grupo de Estudo, verificar o possível efeito da interferência do corretor.

Como dito anteriormente, essa recorreção visa verificarmos se a hipótese de interferência dos corretores se manifesta no PAEBES-Alfa. Para isso, os itens do grupo de Estudo foram corrigidos obedecendo exatamente os mesmos critérios, chave de correção, anteriormente estabelecidos. Realizaram a recorreção dois corretores com experiência em correção de itens de resposta graduada, responsáveis, em diferentes momentos, pela capacitação de corretores de itens de produção escrita de alunos que se encontram no ciclo de alfabetização. É importante destacar que, durante a recorreção, os avaliadores não tiveram acesso às notas recebidas pelos alunos na correção oficial do PAEBES-Alfa.

[1]Conforme Eckes (2011), se a prova/item anterior for muito boa, há uma tendência de subestimar a prova/item seguinte e vice-versa.

Realizamos, para esta pesquisa, a recorreção dos testes do grupo de Estudo, do início e do final do 1º ano, em 2011, e o final do 2º ano, em 2012. Foram realizadas análises estatísticas para cada uma das ondas de avaliação, mas só foi possível comparar o comportamento dos corretores com as informações da correção do final do 2º ano de 2012, uma vez que apenas a partir desse ano tornou-se possível relacionar o corretor ao aluno. Acrescente-se, ainda, que é neste ano de avaliação que os alunos apresentam uma queda na proficiência em escrita.

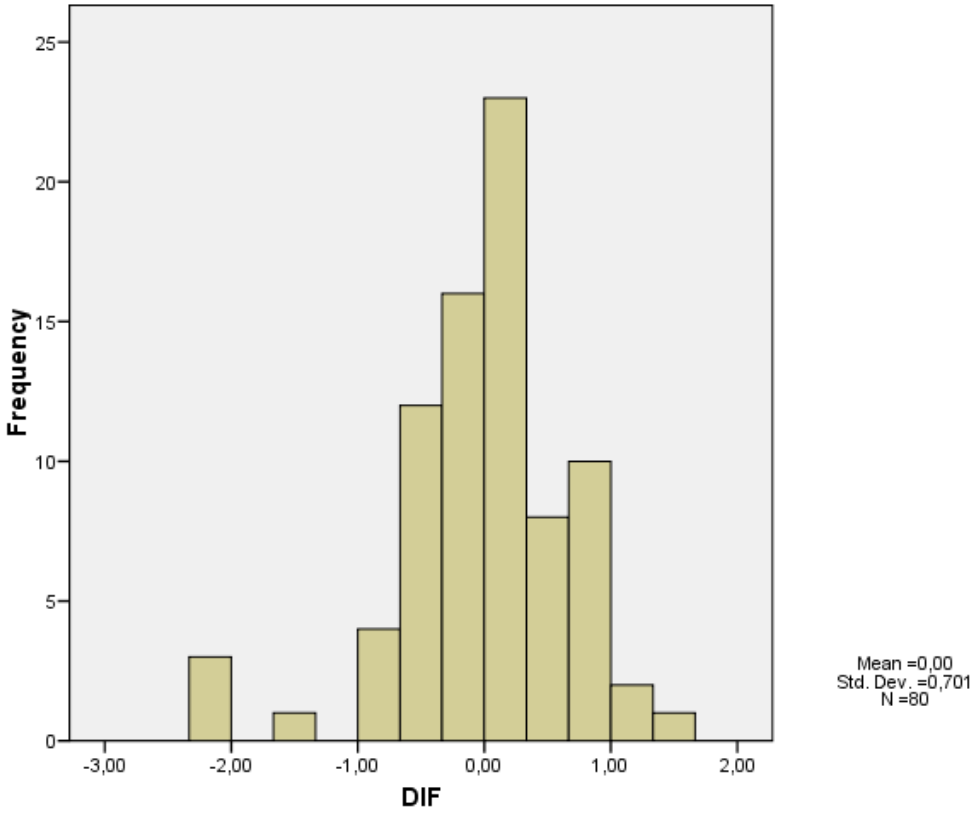
Diante disso, o primeiro passo foi examinar se havia, no grupo de Estudo, alunos que, ao final do 2º ano, foram levados para essa amostra pela concentração de um conjunto de poucos corretores. Esse procedimento visou verificar se a amostra não traria um viés por uma composição de alunos cujas produções tivessem sido submetidas aos mesmos corretores, com características de alta severidade.

Para isso calculamos o percentual de testes corrigidos no grupo Outros e no grupo de Estudo, para cada um dos 82 (oitenta e dois) corretores que compuseram a equipe de correção do PAEBES-Alfa de 2012. Constatamos, com esse procedimento, a presença, no grupo de Estudo, de 98% dos corretores que participaram da correção oficial, garantindo a cobertura de diferentes estilos de correção: severa, leniente e alo. Além disso, essa análise possibilitou-nos observar quais corretores levaram mais ou menos alunos para a nossa amostra.

O Gráfico 6 apresenta a diferença[2] de percentual de testes corrigidos em Outros e testes corrigidos no grupo de Estudo.

[2] DIFERENÇA= percentual corrigido em Outros – percentual corrigido em Estudo.

Gráfico 6: Diferença entre o percentual de testes corrigidos em Outros e o percentual corrigido em Estudo para o 2º ano de 2012.



A análise desse gráfico nos mostra que, à esquerda de 0, posicionaram-se os corretores que corrigiram poucos testes no grupo Outros, mas, no grupo Estudo, encontramos muitos alunos que passaram por sua correção, no 2º ano Ensino Fundamental. Esse comportamento pode ser um indício de corretores mais severos.

Já, à direita, encontram-se corretores que, proporcionalmente, corrigiram muitos textos, mas que levaram poucos alunos para o grupo Estudo.

Encontramos, assim, corretores que, proporcionalmente, jogaram mais casos ou menos casos no grupo de Estudo, conforme sua participação na correção oficial.

4.3
Resultados da recorreção

A possibilidade de comparar a correção oficial com a recorreção realizada para este trabalho pode ser considerada uma maneira de analisar a criticidade do corretor, não da forma como aponta Eckes (2011), pois não

trabalhamos com a modelagem estatística – MRFM- de multifacetas. Contudo, esse procedimento nos leva a observar e analisar o comportamento dos corretores em relação ao grupo e em relação a eles mesmos, permitindo-nos traçar um perfil de correção.

O quadro a seguir traz essa comparação item a item para cada onda de avaliação, indicando o percentual da convergência (ponto 0) ou de discrepâncias, para mais ou para menos, que utiliza a seguinte equação: DIFERENÇA= CORREÇÃO OFICIAL – RECORREÇÃO.

A letra P seguida de um número indica a posição do item considerado na avaliação, e que corresponde a uma certa habilidade indicada na matriz de referência. Nos testes do PAEBES-Alfa, as posições em escrita são ocupadas, sempre, por itens que avaliam a mesma habilidade.

O ponto 0 indica a inexistência de diferença na atribuição das notas. Os pontos acima indicam o quanto a nota da recorreção foi maior do que a nota da correção oficial.

Os valores negativos da DIFERENÇA, acima do ponto 0, indicam que os avaliadores foram mais lenientes na atribuição da nota dos alunos que responderam a determinado item, atribuindo nota mais alta.. Ou seja, interpretaram de modo mais tolerante a produção dos alunos, mesmo de posse da chave de correção, instrumento que visa orientar os avaliadores e padronizar o processo.

Já os valores que se posicionam abaixo do ponto 0 revelam que a nota da recorreção foi menor do que nota indicada na correção oficial. Desse modo, os valores positivos da DIFERENÇA revelam que os avaliadores da recorreção foram mais severos do que os avaliadores da correção oficial para os mesmos itens.

Quadro 13: Comparação entre recorreção e correção oficial por onda de avaliação.

DIF	2011.1					2011.2						2012.2					
	P21	P22	P23	P24.1	P24.2	P21	P22	P23	P24	P25.1	P25.2	P21	P22	P23	P24	P25.1	P25.2
-4	0,6	0,3	0,3	1,1	1,1	0,2	0,2	0,2	0,0	0,2	0,2	0,0	0,2	1,8	0,2	1,4	2,7
-3	1,1	0,8	0,5	0,6	1,0	0,6	0,0	0,3	0,2	2,6	1,0	0,2	0,3	0,8	1,0	4,0	2,4
-2	14,1	6,4	1,9	1,8	1,8	2,7	0,0	0,5	1,1	4,6	1,8	5,9	1,6	3,5	6,2	9,9	8,1
-1	5,8	1,0	3,0	5,1	2,6	1,6	1,3	3,8	3,7	7,8	4,2	10,2	10,1	10,5	21,2	14,1	14,1
0	72,5	88,5	91,5	88,2	77,3	84,7	95,8	88,8	86,1	80,0	80,2	71,9	81,8	78,3	60,4	51,9	49,2
1	2,7	1,9	1,8	1,9	10,9	2,4	1,6	4,8	6,9	3,0	7,7	6,9	4,6	4,0	9,6	14,7	17,1
2	2,9	0,0	0,2	0,5	2,2	6,9	0,3	0,6	0,5	0,6	2,7	4,8	1,1	0,5	1,4	3,0	5,8
3	0,3	0,3	0,0	0,0	1,9	0,6	0,2	0,3	1,0	0,5	2,1	0,0	0,2	0,2	0,0	0,0	0,5
4	0,0	0,2	0,3	0,3	0,3	0,0	0,3	0,0	0,2	0,2	0,0	0,2	0,2	0,5	0,0	1,0	0,2

Analisando esse quadro, observamos um altíssimo índice de convergência entre a recorreção e a correção oficial, o que destacamos em cinza. Para 2011.1 e 2011.2, observa-se, uma média aproximada de 85% de igualdade entre as notas atribuídas na recorreção e na correção oficial em todos os itens.

No entanto chama a atenção, em 2011.1, um item que poderia ser considerado o mais simples (P21) e que solicita ao aluno apenas a cópia de uma frase para verificar se ele domina os princípios do uso da página. No entanto, este item apresenta menor percentual de igualdade em relação à correção oficial, apresentando, ainda, em aproximadamente 15% dos casos corrigidos uma nota maior em dois pontos que a nota atribuída na correção oficial.

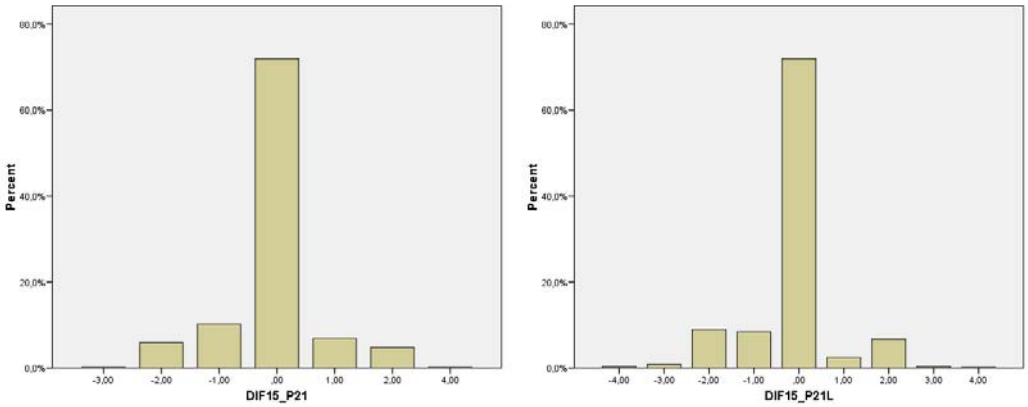
Essa diferença, conforme abordado no Capítulo IV, pode ser decorrente da descrição dos níveis de gradação presentes na chave de correção, que apresenta mescla de elementos a serem avaliados, como a exatidão na cópia e a ocupação do espaço que reproduz uma parte da folha de um caderno. Isso pode sugerir a separação de dimensões a serem consideradas nesse tipo de item ou a incorporação da dimensão “uso da página” em outro item, por exemplo, de escrita de frase a partir de uma cena.

Em 2011.2 e 2012.2, os itens que apresentam o menor índice de convergência são aqueles relacionados à habilidade de produzir textos, os quais se desdobram em duas dimensões, conforme visto no capítulo anterior.

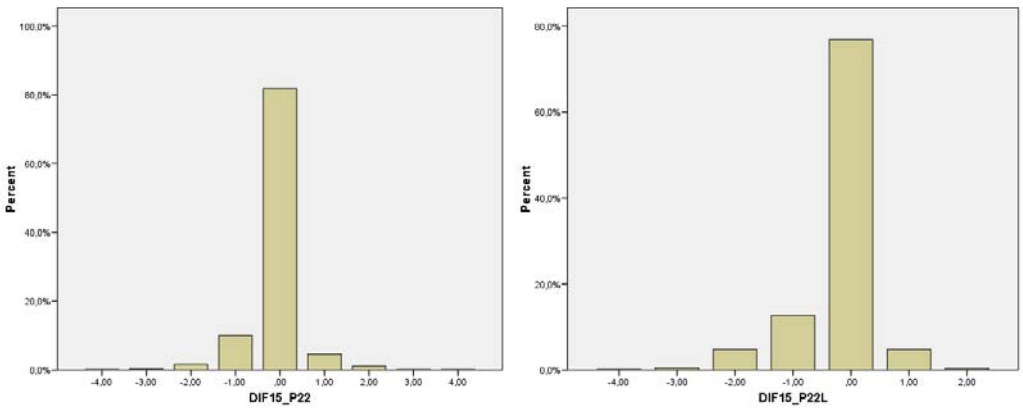
Destacamos que esse baixo índice de convergência ocorre nas duas dimensões dos itens de produção de texto indicado como P25.1 e P25.2.

Os Gráficos de 7 a 12 permitem visualizar de outra forma a recorreção do 2º ano (2012.2). De acordo com esses gráficos, as notas atribuídas pela recorreção convergem, na maioria dos casos, para as notas dadas pelos avaliadores da correção oficial.

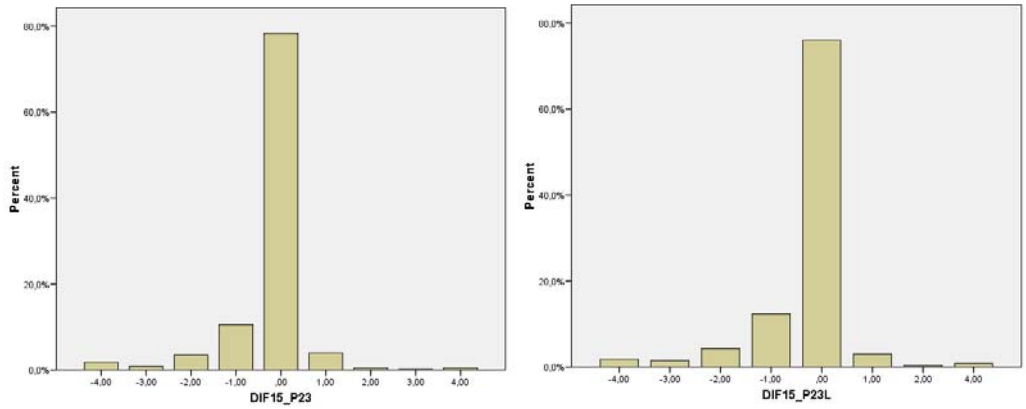
Gráficos 7 e 8: Comparação de correção de uso da página - 2012.2.



Gráficos 9 e 10: Comparação de correção para escrita de palavra com apoio de imagem - 2012.2.



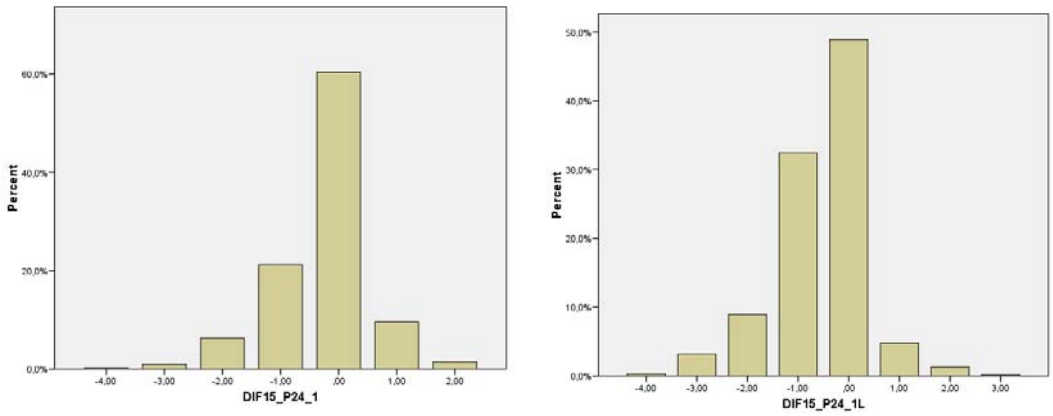
Gráficos 11 e 12: Comparação de correção para ditado de palavra - 2012.2.



Nessa sequência de seis gráficos, para comparação de correção de itens que avaliam o uso da página, escrita de palavras a partir de ditado e escrita de palavras com apoio de imagem, observamos a convergência de atribuição de notas em, aproximadamente 75% dos casos.

No entanto, nos itens que contemplam habilidades um pouco mais complexas, o índice de convergência diminui significativamente, como pode ser observado na sequência de gráficos.

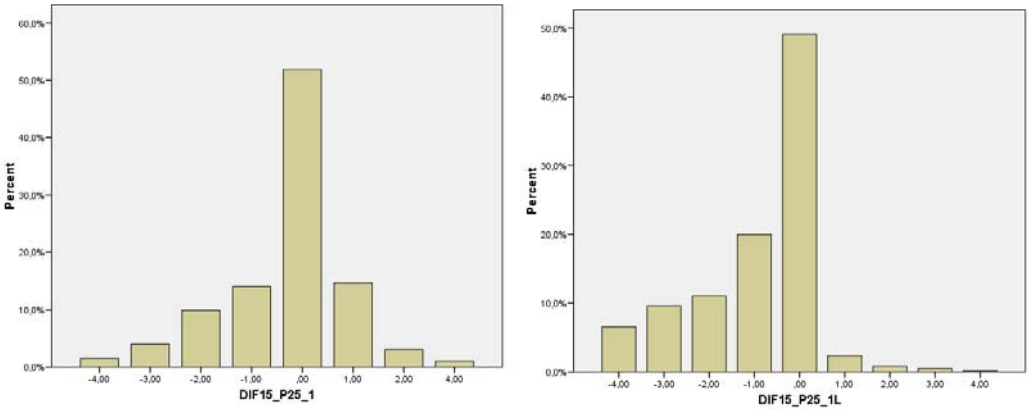
Gráficos 13 e 14: Comparação de correção para escrita de frase (ditada ou com o apoio de imagem) - 2012.2.



Os Gráficos 13 e 14 dizem respeito a itens sobre a escrita de frase (ditada ou a partir de uma cena). Observamos, nesses casos, menores índices de convergência: 60% dos casos, para o corretor 1, e de, aproximadamente, 50% dos itens corrigidos para o corretor 2.

Para os itens que avaliam a produção de texto, a chave de correção se desdobra em duas dimensões (Dimensão 1: gênero/tipo/tema e Dimensão 2: ortografia), ou seja, os item que no teste ocuparam a posição 25 desdobram-se em dois itens. Isto é mostrado pelos Gráficos 15 a 18.

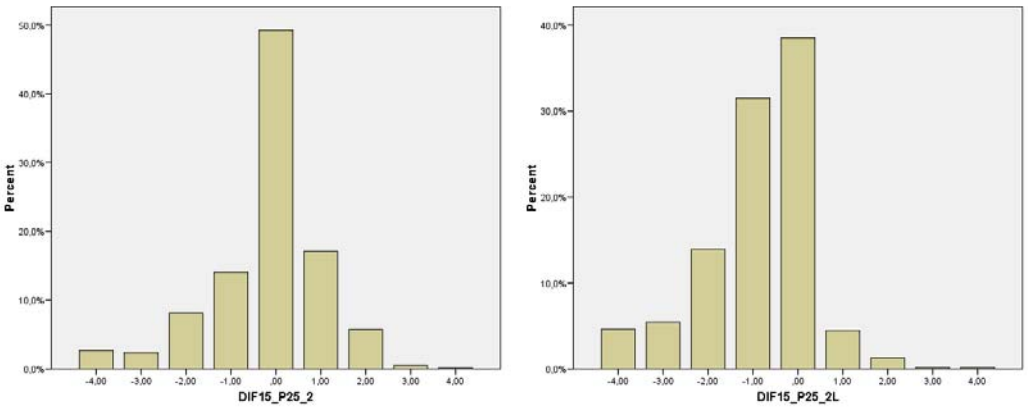
Gráficos 15 e 16: Comparação de correção para produção de lista e de história com apoio de imagem – Dimensão 1 - 2012.2.



Nos itens que propõem a tarefa de produzir uma lista a partir do estímulo de uma imagem e da escuta de uma sequência de palavras de elementos constitutivos da cena, para dimensão 1, observamos que, o corretor 1 apresenta uma convergência em, aproximadamente, 55% dos casos avaliados, enquanto que, para o corretor 2, há 45% de situações de convergência.

Situação semelhante ocorre na avaliação da dimensão relacionada aos aspectos ortográficos, como mostram os Gráficos 17 e 18.

Gráficos 17 e 18: Comparação de correção para produção de lista e de história com apoio de imagem – Dimensão 2 - 2012.2.



Na dimensão 2 dos itens de produção de texto, a atribuição de notas da recorreção só encontra convergência em aproximadamente 45% (corretor 1) e 40% (corretor 2).

Como dito anteriormente, observa-se um equilíbrio na atribuição das notas na comparação realizada entre a recorreção e a correção oficial, principalmente nos itens considerados mais fáceis, conforme o parâmetro *b*.

Apesar de haver uma maior convergência, observamos, na atribuição das notas na recorreção, que o corretor 2 tem uma tendência a ser mais leniente, o que se confirma nos itens de produção de texto.

Essa tendência à leniência se revela na posição das notas por ele atribuídas que, depois de convergirem para a correção oficial, na maioria dos casos se encontram à esquerda do ponto 0. Em outras palavras, esse posicionamento indica que, em todos os itens, quando não converge para a correção oficial, o corretor 2 atribui notas mais altas, fazendo, assim, que a DIFERENÇA observada entre as correções seja negativa.

Situação anunciada nos itens que avaliam o uso da página, escrita de palavras a partir de ditado e escrita de palavras com apoio de imagem, ela se concretiza de modo bastante explícito nos itens de produção de texto (lista e produção de uma narrativa), como podemos observar nos Gráficos 16. e 18, apresentados anteriormente.

No entanto, o fato de termos observado a predominância de uma convergência da recorreção para a correção oficial, permite-nos dizer que ambas as correções mantêm uma relação mas não nos informa sobre o grau de associação entre essas duas correções. De acordo com Levin e Fox (2004), a diferença na intensidade da correlação pode ser visualizada por meio de gráficos de dispersão.

Assim, para visualizarmos a intensidade da correlação existente entre os dois processos de correção apresentamos os Gráficos de 18 a 21para cada uma das ondas de avaliação submetidas à recorreção.

Gráfico 19: Correlação entre correção oficial e recorreção 2011.1

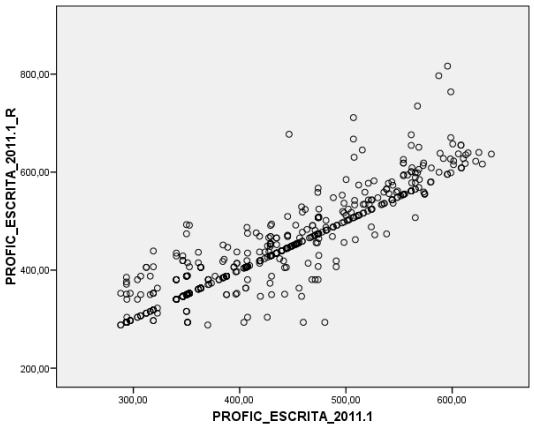


Gráfico 20: Correlação entre correção oficial e recorreção 2011.2

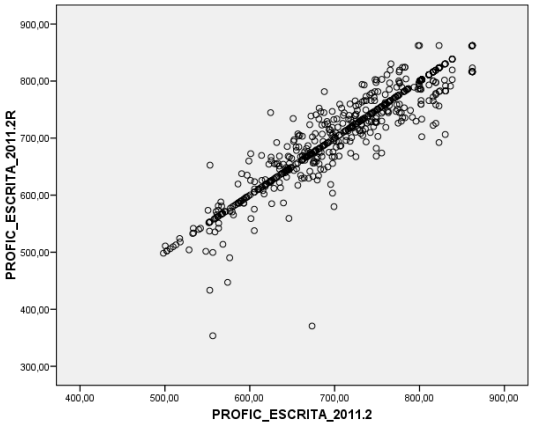
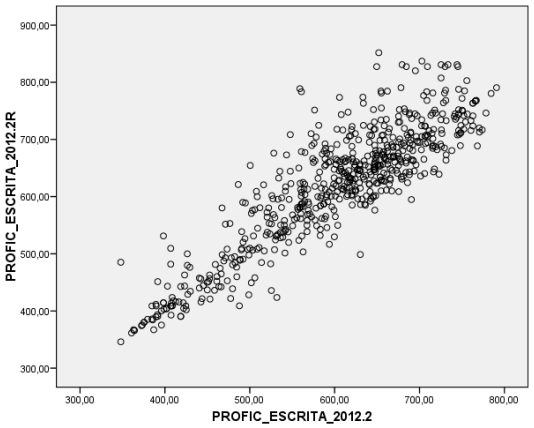


Gráfico 21: Correlação entre correção oficial e recorreção 2012.2



Ao analisarmos os três gráficos constatamos que a correção oficial e as recorreções mantêm uma correlação muito forte, pois percebemos, nos três casos, uma aglomeração de em torno de uma reta diagonal imaginária, conforme apontam Levin e Fox (2004). Em termos de coeficiente de correlação, na situação do gráfico xxx (início do 1º ano), o R^2, o coeficiente de correlação é de 84% e sobe para 86,5% ao final do 2º ano. Já para o final do 2º ano, a correlação permanece forte, mas se observa uma ligeira queda no coeficiente de correlação, agora de 70%. Essa variação pode ser decorrente, por exemplo, do perfil dos corretores ou da característica do teste ou do aluno.

Considerando a forte correlação entre os dois processos de correção, na nova análise do desempenho dos alunos após a recorreção, observamos que 90% dos 626 alunos do grupo de Estudo continuam a apresentar um aumento significativo de proficiência ao final do 1º ano e uma queda acentuada ao final do 2º ano, mesmo se não considerarmos o corte estabelecido para a constituição do grupo de Estudo, ou seja, aumento de 200 e queda de 50 pontos na proficiência.

Quando tomamos os resultados da recorreção, considerando o perfil de avanço de 200 pontos na escala de proficiência, ao final do 1º ano, e, a queda de 50 pontos ou mais, ao final do 2º ano, temos 61,5% dos casos em que os alunos mantêm exatamente o mesmo desempenho apresentado com a correção oficial.

O alto grau de correlação observado entre a correção oficial e a recorreção, assim como os resultados observados com a recorreção, permitem constatar que, realmente, há corretores com perfis diferentes, mas que essa dimensão individual, do ponto de vista estatístico, não afeta o resultado global da avaliação do PAEBES-Alfa. Assim, desconsideramos a hipótese do efeito-corretor sobre os resultados do grupo Estudo e passamos a reconsiderar as outras duas hipóteses formuladas para explicar esses resultados, relativas à dificuldade dos testes do final do 1º e do 2º ano e à questão das condições de aplicação dos testes.

Para a análise da dificuldade dos testes, tomamos 61,5 % dos alunos do grupo Estudo, composto de 385 (trezentos e oitenta e cinco) alunos, a partir de agora, denominado de grupo Estudo Recorreção (EstudoR). Esses alunos

têm em seus resultados o filtro do efeito da correção: a partir da recorreção, para esse grupo de alunos, não se observou qualquer alteração nos resultados.

Assim, uma vez que, para grupo EstudoR, não há qualquer interferência da correção na produção de seus resultados, tomamos esse grupo e o comparamos ao grupo Outros, para poder realizar uma análise dos testes, em função de seu ajuste à população avaliada. Associando essa análise a uma interpretação de desempenho por meio dos padrões de desempenho em escrita do PAEBES-Alfa, buscamos verificar, ainda, a complexidade do teste por meio da análise dos pontos de ancoragem dos itens de escrita.

Considerando o grupo EstudoR, ao filtrarmos o quantitativo de alunos por escola, encontramos alguns casos que chamam a atenção pela concentração de alunos com essa característica de apresentarem, ao final do 1º ano, um aumento de 200 pontos na escala de proficiência e uma queda de 50, ao final do 2º ano.

Consideramos, para isso, as escolas que levaram mais de quatro alunos para o grupo pesquisado. Encontramos, assim, escolas pertencentes às redes municipais, à rede estadual e a escolas particulares.

O Quadro 14 mostra as escolas, os municípios onde elas se localizam, a rede a que pertencem, o número de alunos que participaram das três ondas de avaliação, assim como o número de alunos que se encontram no grupo EstudoR e o que isso significa em termos de percentuais.

Quadro 14: Escolas com mais de quatro alunos do EstudoR.

MUNICÍPIO	ESCOLA	REDE	QUANTIDADE DE ALUNOS		PERCENTUAL (%)
			TOTAL	ESTUDOR	
MUNIZ FREIRE	EEEFM ARQUIMIMO MATTOS	ESTADUAL	26	13	50
SERRA	EEEFM GETULIO PIMENTEL LOUREIRO	ESTADUAL	16	7	44
CACHOEIRO DE ITAPEMIRIM	EMEB CORREGO VERMELHO	MUNICIPAL	11	4	36
VILA VALERIO	EMEF MARIA LUIZA JORGE DOS REIS	MUNICIPAL	11	4	36
MONTANHA	EMEF PEDRO PALACIOS	MUNICIPAL	14	5	36
SAO MATEUS	EEEFM NESTOR GOMES	ESTADUAL	23	8	35
MARATAIZES	EMEF PROF LAUREA FREIRE BRUMANA	MUNICIPAL	27	9	33
SERRA	EEEF PROF ANNA GOMES	ESTADUAL	36	8	22
CACHOEIRO DE ITAPEMIRIM	EEEF ROTARY	ESTADUAL	55	10	18
MARATAIZES	EMEF MARIA DA GLORIA NUNES NEMER	MUNICIPAL	30	4	13
ALEGRE	EEEF PROFESSOR LELLIS	ESTADUAL	38	5	13
CACHOEIRO DE ITAPEMIRIM	CE SAO CAMILO - ICE	PARTICULAR	35	4	11
NOVA VENECIA	EMEF TITO DOS SANTOS NEVES	MUNICIPAL	43	4	09
PEDRO CANARIO	EMEF PROF GUEDES ALCOFORADO	MUNICIPAL	46	4	09
CACHOEIRO DE ITAPEMIRIM	EEEF INAH WERNECK	ESTADUAL	52	4	08
SERRA	EMEF NOVO HORIZONTE	MUNICIPAL	76	4	05
VILA VELHA	UMEF LEONEL DE MOURA BRIZOLA	MUNICIPAL	112	4	04

Essas escolas, assim, como aquelas que apresentam um único aluno que se encontra no grupo EstudoR deveriam ser objeto de uma pesquisa qualitativa que permitisse verificar quais fatores associados às práticas de ensino, ou mesmo à situação de aplicação dos testes, podem ter levado a esse desempenho.

Neste estudo, no entanto, as próximas seções se debruçam sobre a análise dos testes.

4.4
A análise do ajuste dos testes por categoria de resposta

Uma análise importante para se observar a adequação do teste à população avaliada é denominada de ajuste da curva do teste. Isso é possível, pois a TRI nos permite colocar em uma mesma métrica a dificuldade do item e a proficiência dos alunos.

Os gráficos de ajuste do item à população foram construídos calculando-se o percentual de aluno por faixa de desempenho e o percentual de itens, nessas mesmas faixas. O item foi posicionado na faixa através de seu ponto de ancoragem, que corresponde à proficiência que o aluno deve ter para que ele tenha 65% possibilidade de acertar esse item.

O modo como essa análise vinha sendo feita, até o momento, considerava o ponto de ancoragem de cada categoria de resposta, entendendo, cada uma delas como sendo um item. Ao analisarmos as curvas de ajuste do teste para o grupo Outros, concebidas dessa maneira, constatamos que o teste do início do 1º ano é um pouco difícil para o público avaliado, ou seja, os itens se encontravam mais à esquerda da escala enquanto os alunos se posicionaram mais à direita. Isso se explicaria pelo fato de tratarmos de testes aplicados a alunos que acabaram de ingressar no Ensino Fundamental. Já para avaliações do final do 1º e do 2º ano, observamos um melhor ajuste do teste à população: os itens avançaram para a esquerda, e a população acompanhou esse avanço.

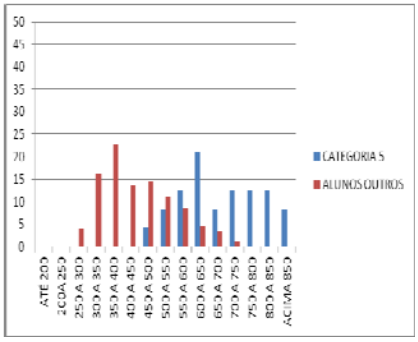
Em particular, para este trabalho, nos propusemos a construir curvas de ajuste do teste para cada categoria de resposta indicada na chave de correção, da mais correta para a menos correta. Temos, assim, a seguinte correspondência entre gradação/categoria da chave de correção e notas/categorias para geração de resultados: A=5; B=4; C=3 e D=4. Como o modelo é graduado, parte-se do princípio de que todos os alunos têm a categoria E, a mais básica, daí a sua ausência nas análises.

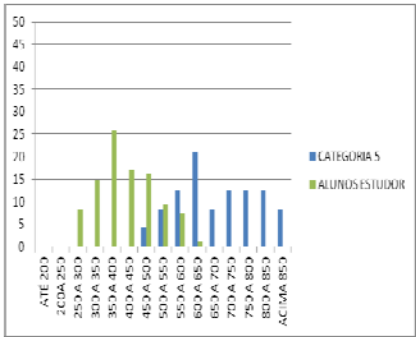
Essa proposta de análise de ajuste por categoria de resposta, repousa na possibilidade de obtermos dados mais precisos acerca da dificuldade do teste, assim como informações mais detalhadas relacionadas ao processo de aprendizagem da escrita.

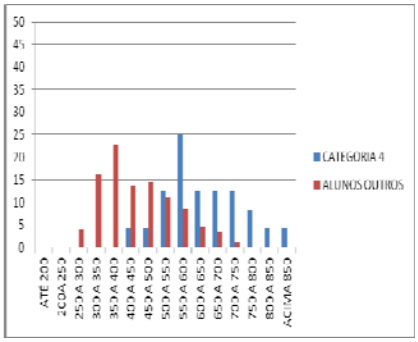
Assim, ao trabalharmos com essa proposta de análise de ajuste ao teste, observamos um comportamento diferente do teste em relação à população. Os Gráficos de 22 a 29 apresentam esse comportamento.

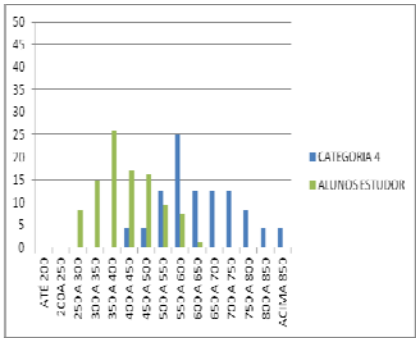
Esses gráficos apresentam, no eixo X, os intervalos da escala de proficiência e, em Y, a indicação de percentual de itens/alunos que se encontram em determinada faixa. Em todos os gráficos, o azul representa os itens que entraram no teste, o vermelho, os alunos do grupo Outros e o verde, os alunos do grupo EstudoR.

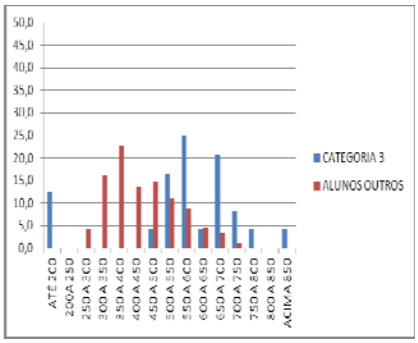
Gráficos 22 a 29: Curvas de ajuste do teste à população 2011.1.

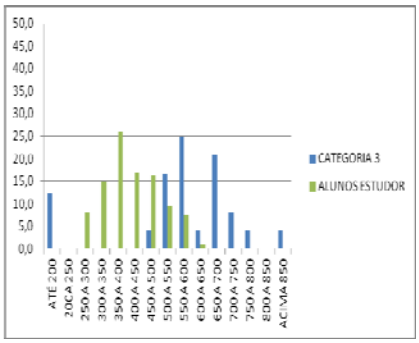


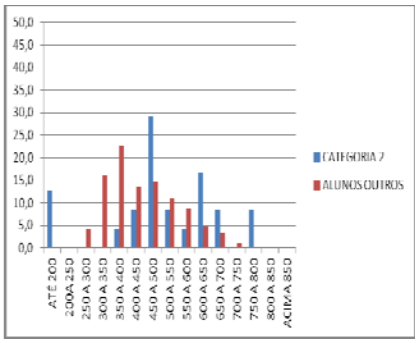


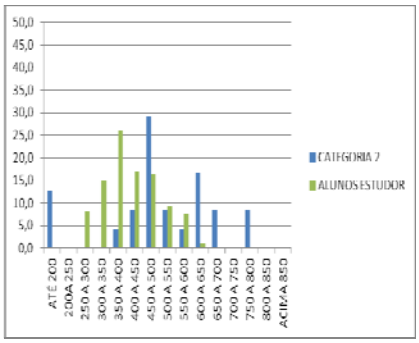












Ao analisarmos os gráficos de ajuste do teste para os Outros e para o EstudoR, considerando, separadamente, cada uma das categorias de resposta, deparando-nos com uma situação um pouco diferente daquela observada na curva de ajuste quando tomamos o conjunto de categorias.

Comparando os dois grupos, constatamos que, para o grupo EstudoR, o teste foi ainda mais difícil, em todas a categorias, pois há mais alunos concentrados nas faixas de 200 a 250 até a faixa de 450 a 500, sem que encontremos qualquer aluno posicionado acima de 650 pontos na escala.

Com relação ao ajuste do teste à população, observamos, para os dois grupos, uma tendência de ajuste que não se revela totalmente perfeita. Tal tendência se apresenta na categoria 2 (D), para os dois grupos, sendo mais evidente para o grupo Outros que para o grupo EstudoR, pois como vimos este grupo apresenta uma proficiência média bem baixa.

Como dito anteriormente, esse procedimento de análise de ajuste por categoria de resposta, permite-nos observar mais claramente a etapa do processo de aprendizagem da escrita em que o aluno se encontra, e que pode ser, como discutido no Capítulo I, associada ao modelo de etapas de Frith e às hipóteses de Ferreiro e Teberosky.

Assim, percebemos que, para os itens que consideram aspectos associados à aprendizagem da ortografia, os alunos se encontram na etapa denominada por Frith de alfabética e de silábica ou silábico-alfabética por Ferreiro e Teberosky. Isto é, os alunos ainda trocam e espelham letras, apresentam problemas tanto de hipo quanto de hipersegmentação de palavras ou, ainda, escrevem ora com uma letra representando um som, ora com uma letra representando uma sílaba.

Para os itens que avaliam a produção de texto, na dimensão 1, associada aos aspectos de gênero, tipologia, tema e coesão/coerência, esses alunos conseguem, por exemplo, na situação de produção de uma narrativa com estímulo de imagem, escrever uma história que tem pouca relação com a cena utilizada como estímulo, sem que haja, necessariamente, a utilização de conectivos.

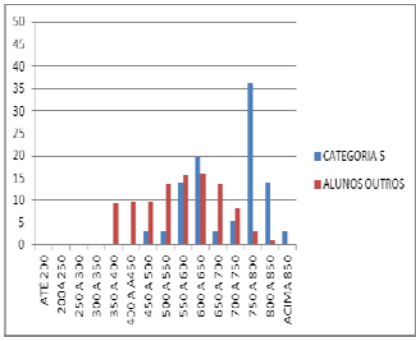
No caso da produção de uma lista, conforme a chave de correção, esse alunos escrevem “uma ou duas palavras sugeridas pela, mas que não foram

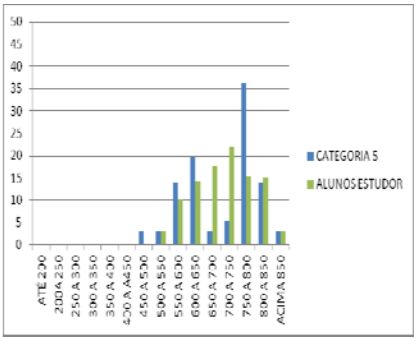
solicitadas pelo enunciado". Estamos, pois, diante da situação relacionada a itens que apresentam dois estímulos (visual/auditivo), na qual o aluno acaba sendo penalizado não por ter cometido desvios ortográficos ou de adequação ao gênero, mas pela mescla dos estímulos que envolvem o item.

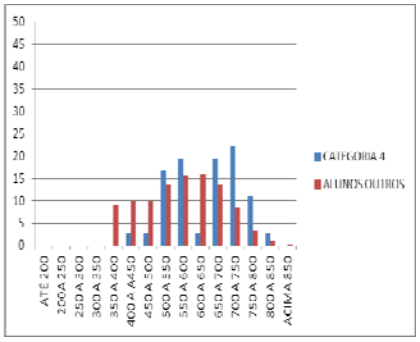
Em face do exposto, constatamos que os alunos dos dois grupos encontram-se, ainda, em um processo bastante inicial de aprendizagem da língua escrita.

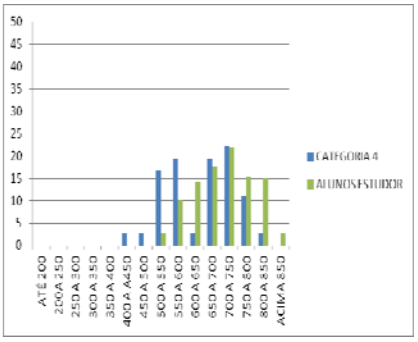
No entanto, ao final desse mesmo ano de escolaridade (2011.2), esses mesmos alunos avançam significativamente, e este avanço pode ser melhor visualizado após análise das curvas de ajuste do teste, nos Gráficos de 29 a 36.

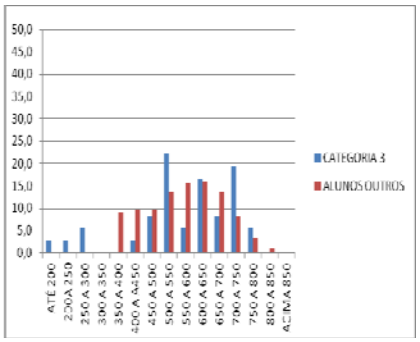
Gráficos 30 a 37: Curvas de ajuste do teste à população 2011.2

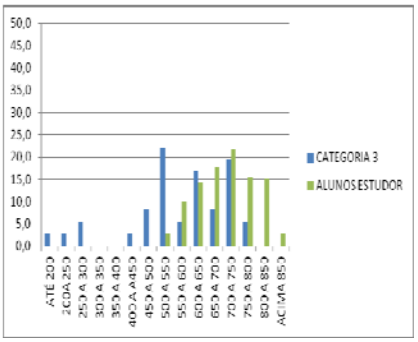


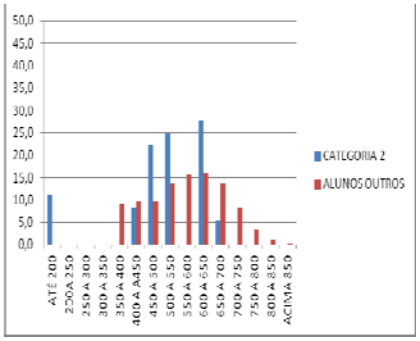


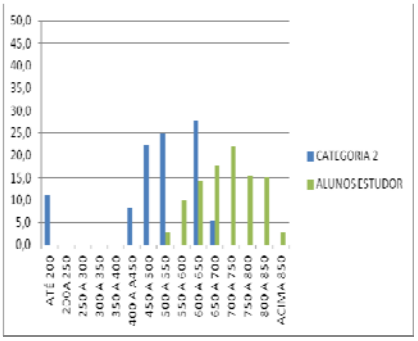












Comparando os dois grupos, considerando os Gráficos de 29 a 36, notamos que, no teste do final do 1º ano, há diferença no avanço da proficiência do grupo Outros em relação ao grupo EstudoR. No grupo Outros, os alunos se

distribuem em torno da média da escala (500 pontos), enquanto os alunos de EstudoR deslocam-se para a direita, alocando-se acima da faixa de 500 a 550, num comportamento oposto ao que foi observado no início do ano. No entanto, o teste mantém, em cada categoria, as mesmas características de dificuldade, indicando, assim, que é um teste bastante difícil.

No que diz respeito ao ajuste desse teste à população, observamos que, nessa onda de avaliação, há uma diferença na categoria na qual ocorre esse ajuste.

Para o grupo Outros, o teste se ajusta na categoria 3 (C), que pode se considerada uma etapa intermediária do processo de consolidação das habilidades. Nessa categoria, considerando o modelo de Frith, os alunos se encontram, ainda, em uma etapa ortográfica e, para Ferreiro e Teberosky, eles apresentam uma hipótese alfabética ou mesmo silábico-alfabética.

Assim os alunos do grupo Outros, nos itens de escrita de palavras e de frases, com relação à ortografia, revelaram que ainda podem apresentar letras espelhadas e alguns desvios relativos à segmentação das palavras, podendo acrescentar ou omitir letras, principalmente nos dígrafos.

Com relação aos itens de produção de texto, esses alunos, na dimensão 1, revelaram-se capazes de produzir um texto coerente com a proposta apresentada, ainda que faltem alguns elementos estruturadores do gênero solicitado. Começam também a utilizar, em narrativas, marcadores de progressão textual que trazem fortes interferências da oralidade.

Em contrapartida, os alunos do grupo EstudoR, como visto anteriormente, avançam de modo acentuado na escala (sobem 200 pontos), e o teste se ajusta em uma categoria acima daquela do grupo Outros. Para o grupo EstudoR, o teste se ajusta na categoria 4 (B), para todos os itens. Isso significa que esses alunos já se encontram em uma etapa ortográfica, seguindo o modelo de Frith, e alfabética, de acordo com as hipóteses de Ferreiro e Teberosky.

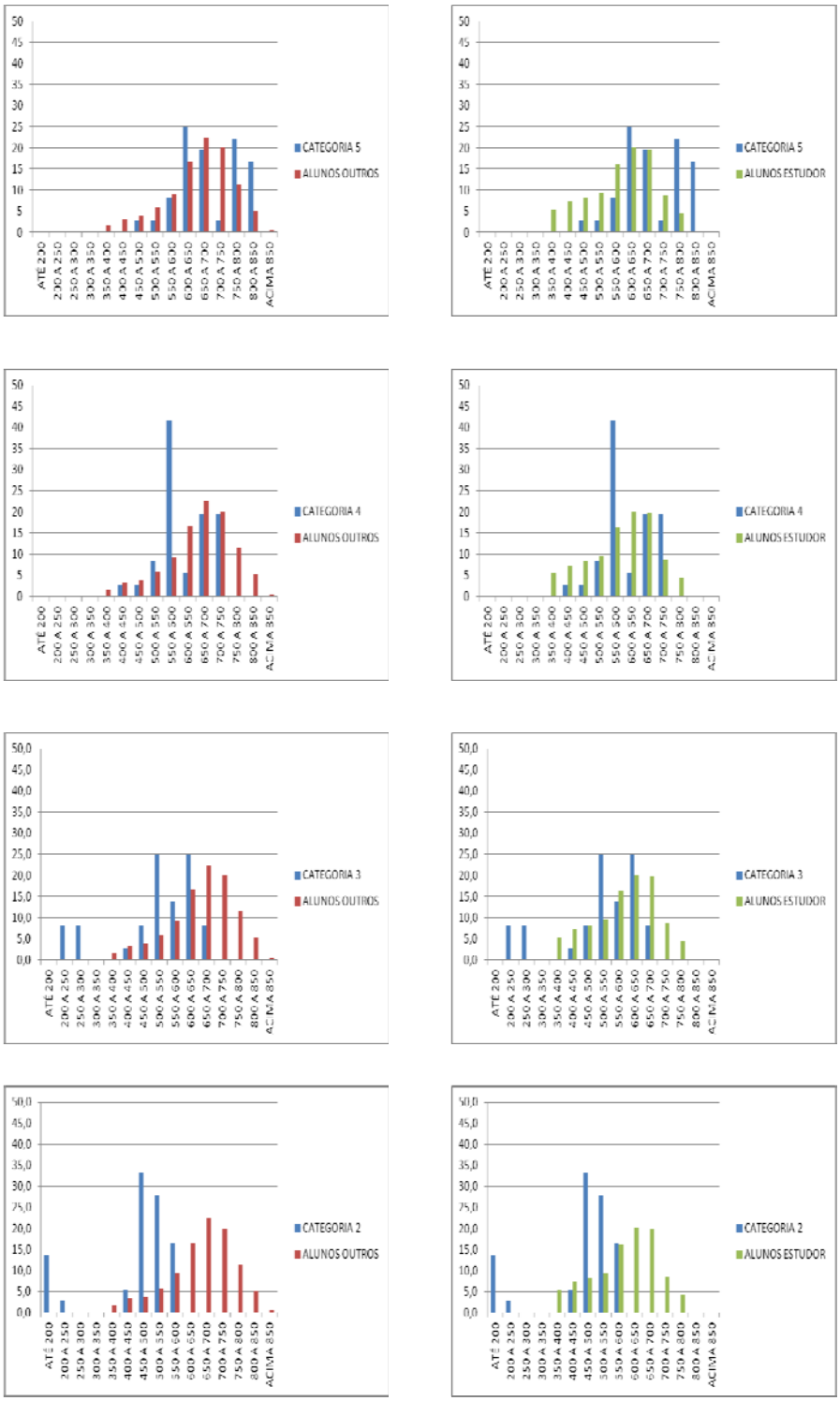
Ao se posicionarem nessa categoria, os alunos EstudoR cometem desvios de ortografia relacionados, principalmente, aos dígrafos e às representações dos fonemas [z] e [s]. Eles já conseguem, na produção de textos, atender à proposta solicitada, omitindo um elemento constitutivo do gênero solicitado, como, por exemplo, em um bilhete, não indicar a data, ou em

um relato, não contemplar o elemento “tempo”, além de começarem a realizar operações de substituição por meio do uso de substantivos.

Diante do exposto, constatamos que o avanço do grupo Outros parece ser mais coerente com o processo de aprendizagem da escrita que o do grupo EstudoR, cujos alunos demonstraram o desenvolvimento de habilidades ainda mais elementares do que Outros no teste do início do 1º ano. Nesse sentido, o comportamento de desempenho dos alunos de EstudoR destoa como um resultado não esperado do ponto de vista cognitivo, principalmente se considerarmos que anulamos a possibilidade da interferência do efeito corretor sobre este grupo.

Quando acompanhamos esses grupos até o final do 2º ano (2012.2), se configura a situação que pode ser visualizada nos Gráficos 37 a 44 para o ajuste do teste à população.

Gráficos 38 a 45: Curvas de ajuste do teste à população 2012.2.



Analisando a distribuição das categorias ao longo da escala, observamos que, no teste de 2012.2, houve um deslocamento dos itens para a

esquerda na escala, o que é um indício de que o teste dessa onda de avaliação foi mais fácil do que o teste do ano anterior.

Com relação ao posicionamento dos alunos do grupo Outros, esse gráfico mostra que eles avançam significativamente na escala, o que se concretiza quando, ao analisarmos o ajuste do teste à população, nos deparamos com o ajuste ocorrendo na categoria 5 (A), que exige o desenvolvimento de habilidades mais elaboradas.

O avanço dos alunos do grupo Outros revela que eles se encontram na etapa ortográfica do modelo de Frith e na hipótese alfabética de Ferreiro e Teberosky. Isto é, nos itens que avaliam aspectos ortográficos, esses alunos demonstraram já ter aprendido os princípios que regem o sistema alfabético, não cometendo, portanto, desvios na ortografia.

Para os itens que avaliam a produção de texto, ou seja, de itens que avaliam desde a produção de um bilhete até a elaboração de pequenas narrativas ficcionais ou não, o teste se ajusta na categoria 5 (A). Isto significa que o texto atende à proposta, no que diz respeito ao tema e às características do gênero ou tipologia solicitadas.

Em contrapartida, os alunos do grupo EstudoR comportam-se de modo semelhante ao teste como um todo. Isto é, os itens se deslocam para a esquerda e, também, os alunos do grupo EstudoR.

Com relação aos itens, significa que apresentam uma dificuldade menor, e que alunos com menor proficiência têm 65% de chances de acertá-los. Quanto aos alunos, esse deslocamento para a esquerda, ou seja, para níveis mais baixos na escala, indica uma queda na proficiência.

Assim, esse desempenho faz com o teste se ajuste à população na categoria 3 (C) o que revela, na escala de proficiência, no que diz respeito às habilidades, um retrocesso na direção da capacidade de realizar, no 2º ano, tarefas mais simples, inclusive que aquelas já desenvolvidas ao final do 1º ano.

Embora essa nova forma de analisar o ajuste do teste à população forneça mais informações acerca das habilidades esperadas para um aluno conseguir realizar as tarefas propostas, podemos refinar um pouco mais ainda a nossa análise, chegando ao nível de ajuste do teste à população por meio da ancoragem dos descritores dentro dos padrões de desempenho.

4.5.
Análise do ajuste do teste à população pelos padrões de desempenho

O fato de cada item de escrita apresentar cinco categorias de resposta e da possibilidade de alguns deles serem avaliados com mais de uma chave de correção, acaba gerando uma grande quantidade de parâmetros de dificuldade pela TRI e, consequentemente, uma excelente análise do ajuste do teste à população.

Dessa forma, procuramos analisar esse ajuste levando também em consideração o cruzamento do ponto de ancoragem médio de cada descritor nas diferentes categorias de respostas com os padrões de desempenho dos alunos, o que deu origem ao Quadro 15.

Nesse quadro, representamos, com a cor verde, o ponto de ancoragem, ou seja, o ponto em que a habilidade está em processo de desenvolvimento. Já a cor amarela indica que a habilidade ainda não foi desenvolvida pelo aluno e a cor azul indica a consolidação da habilidade.

Quadro 15: Faixa do ponto de ancoragem por categoria de resposta

DESCRITOR	CATEGORIA	PADRÃO DE DESEMPENHO											
		2011.1				2011.2				2011.2			
		ABAIXO DO BÁSICO	BÁSICO	PROFICIENTE	AVANÇADO	ABAIXO DO BÁSICO	BÁSICO	PROFICIENTE	AVANÇADO	ABAIXO DO BÁSICO	BÁSICO	PROFICIENTE	AVANÇADO
D05	2	x				x				x			
	3	x				x				x			
	4			x				x			x		
	5				x			x				x	
D28	2		x				x			x			
	3			x			x				x		
	4			x				x			x		
	5			x				x				x	
D29	2			x			x			x			
	3				x			x			x		
	4				x				x		x		
	5				x				x			x	
D30.1	2	x							x	x			
	3				x				x			x	
	4				x				x			x	
	5				x				x				x
D30.2	2		x						x	x	x	x	
	3				x				x			x	
	4				x				x				x
	5				x				x				x

Podemos interpretar esse quadro da seguinte forma. Por exemplo, um aluno Avançado na avaliação 2011.1 está na categoria 5 (A) dos descritores D05 e D2, mas, com relação aos descritores D29, D30.1 e D30.2, ele poderá estar nas categorias 3 (C) ou 4 (B) ou 5 (A).

Assim, em função do comportamento de ancoragem desses últimos descritores falta precisão para descrever o desempenho do aluno em relação às habilidades desenvolvidas nesse padrão de desempenho, pois os itens são muito difíceis, ancorando em pontos muito altos na escala.

Considerando o comportamento desses descritores, se tomarmos como exemplos os alunos que, em 2011.2, se encontravam no padrão Avançado, deparamo-nos com uma situação crítica, pois nessa onda de avaliação o descritor D30, avaliado em suas duas dimensões, o D30.1 e o D30.2 apresentaram um nível de dificuldade muito alto para os alunos do grupo EstudoR. Isso ocorre, pois todas as categorias de resposta dos itens estão ancorando no padrão Avançado.

Isso é um pouco problemático, pois, no 1º ano, o padrão de desempenho é composto pelas faixas acima de 600 pontos na escala de proficiência. Assim, sendo, a descrição das habilidades associadas ao padrão Avançado torna-se pouco precisa, visto que, com relação a D30.1 e D30.2, os alunos que posicionam acima de 600 podem se encontrar em qualquer uma das categorias de resposta (A=5; B=4, C=3 ou D=2).

Assim, para uma análise mais precisa, teremos que, além de informar que o aluno está no padrão Avançado, informar, também, o seu nível de proficiência, conforme mostra o Quadro 17.

O Quadro 16 traz a faixa de proficiência, a categoria de resposta e o percentual de alunos do grupo EstudoR, no final do 1º ano.

Quadro 16: Relação entre padrão de desempenho, proficiência e categoria de resposta de D30.

PADRÃO DE DESEMPENHO	NÍVEL	CATEGORIA DE RESPOSTA D30.1 E D30.2	PERCENTUAL DE ALUNOS (%)
AVANÇADO	600 A 650	2	16
	650 A 700	3	14
	700 A 750	4	8,6
	ACIMA DE 750	5	5

Ao analisarmos esse quadro, constatamos que apenas 5% dos alunos de EstudoR se encontram no padrão Avançado e na categoria 5 (A) dos descritores D30.1 e D30.2. Assim, com relação a habilidades muito difíceis, se considerarmos apenas o padrão de desempenho, ficamos com informações que podem ser consideradas vagas acerca do desempenho do aluno e mesmo para a montagem dos testes.

No entanto, em relação à avaliação de 2012.2, já é possível uma melhor interpretação dos resultados dos alunos que se encontram no padrão Avançado, pois para esse ano de escolaridade nas habilidades D30.1 e D30.2, os alunos encontram-se na categoria de resposta 5 ou 4.

Essa situação evidencia que o teste do final do 1º ano (2011.1) é mais difícil do que foi aplicado em 2012.2, em função das características de D30.1 e D30.2. Dito de outra maneira, a habilidade D30 (Produção de texto), nas duas imensões, foi muito mais complexa para os alunos do final do 1º ano do que para os alunos do 2º ano.

Ao voltarmos aos cadernos de testes e analisarmos os itens, constatamos que, para 2011.1, foi solicitada a produção de narrativas (ficcionais ou não, com ou sem apoio de imagem).

Já para 2012.2, solicitou-se a produção de uma lista de elementos presentes em uma cena. Nesse item, o aluno tem dois estímulos (escuta uma sequência de palavras e vê a imagem). No Capítulo IV, ao abordarmos os instrumentos de avaliação percebemos que esse item, pela forma como foi elaborado, acaba por confundir o aluno pelo duplo estímulo, quando, na verdade, solicita apenas a escrita de palavras.

A outra tarefa solicitada para a produção de texto, em 2012.2, foi escrever uma história, a partir do estímulo de uma história em quadrinhos não verbal, o que envolve, além do processamento semântico, aspectos ortográficos e morfossintáticos.

Ao compararmos os tipos de tarefas envolvidas em D30, e levando em conta o período de escolaridade avaliado, percebemos que faltou ao teste uma gradação de complexidade. Entendemos, ainda, que a avaliação da produção de texto deveria incluir mais do que as duas dimensões já apresentadas nas chaves de correção do PAEBES-Alfa, pois existem aspectos importantes relativos à produção do texto, relacionados a aspectos pragmáticos, de coesão e de coerência, conforme aponta Costa Val (1991), que deveriam ser contemplados. Por isso, consideramos importante que haja, nas chaves de correção de produção de texto, o desdobramento para outras dimensões.

Outra constatação importante é a grande variedade de faixas de proficiência em que D28 (Escrita de palavras) tem seus pontos de ancoragem. Por exemplo, na categoria 5 temos itens ancorando em 5 níveis diferentes. Esse comportamento dos itens que avaliam essa habilidade deve-se ao fato de as palavras solicitadas apresentarem diferentes níveis de complexidade, relacionados à estrutura silábica, extensão, tipo de relação entre fonema e grafema, conforme apontam Lemle (1990) e Scliar-Cabral (2003)

Diante do exposto, ressaltamos a importância de se observar a dificuldade dos testes, considerando-se as categorias de resposta (A, B, C e D) separadamente e não em conjunto como vinha sendo feito até o momento. Acrescente-se ainda a observação de uma adequada gradação da complexidade das tarefas que os compõem de modo a garantirmos uma construção de uma medida cada vez mais precisa em testes de avaliação da alfabetização.