Rafael de Araujo Álvares Marinho

O USO DE AVALIAÇÕES ESCOLARES ORDINÁRIAS PARA ESTUDAR A EVOLUÇÃO DA COMPETÊNCIA EM FÍSICA

Belo Horizonte
Faculdade de Educação da UFMG
2010

Rafael de Araujo Álvares Marinho

O USO DE AVALIAÇÕES ESCOLARES ORDINÁRIAS PARA ESTUDAR A EVOLUÇÃO DA COMPETÊNCIA EM FÍSICA

Dissertação apresentada ao Curso de Mestrado da Faculdade de Educação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Educação.

Linha de Pesquisa: Educação em Ciências

Orientador: Prof. Dr. Oto Borges

Belo Horizonte
Faculdade de Educação da UFMG
2010

Dedico este trabalho a meus filhos, Pedro e Joana, e à memória de meu pai.

AGRADECIMENTOS

Sou sinceramente grato a todos que, de alguma forma, contribuíram para que este trabalho pudesse ser realizado.

Agradeço ao meu orientador, Oto Borges, por compartilhar comigo um pouco do seu enorme conhecimento e cultura.

Agradeço à minha esposa Marina; a meus pais, Roberto e Elizabeth; à minha irmã, Mariana; e à Tetê. Os esforços de todas essas pessoas foram absolutamente necessários para que eu pudesse me concentrar no trabalho.

Agradeço aos professores das disciplinas cursadas. Especialmente ao Arnaldo, por ter contribuído para meu crescimento como pesquisador.

Agradeço aos colegas de mestrado e do COLTEC: Dilvana, Morgana, Beth, Geide, Amanda, Cristiano, Terezinha, Tereza, Larissa, Valmária, Matheus, Wanderson, Tuiã, Josimeire, Tarciso, Talim e Helder.

E agradeço a todos os amigos e familiares que, mesmo de longe, sempre torceram pelo meu sucesso.

RESUMO

A abundância das avaliações de sala de aula e sua íntima relação com o currículo real justificam uma investigação das possibilidades de seu uso em pesquisas. O objetivo deste trabalho é investigar algumas possibilidades e limitações do uso de avaliações escolares ordinárias para estudar a evolução da competência em física. São discutidas algumas características das avaliações escolares, suas relações com a competência e algumas possíveis vantagens de seu uso em relação ao uso de testes de pesquisa ou avaliações sistêmicas. Na análise, utilizam-se dois tipos de avaliações: notas trimestrais e respostas a provas fechadas. Para cada um desses tipos, é feita uma análise multinível longitudinal e os resultados são comparados entre si e também com a literatura. Essas comparações indicam que as avaliações escolares podem ser usadas em estudos da evolução da competência. Porém, destacam-se três ressalvas quanto a esse uso: leva a certa vagueza na conceituação de competência; pode apresentar dificuldades para equalização; e pode não ser adequada a um tratamento unidimensional. Todas essas ressalvas nascem de limitações no desenho metodológico impostas pela ética da prática educativa. Por fim, são discutidas algumas limitações do estudo.

Palavras chave: Competência em física. Avaliações escolares. Notas escolares. Estudos longitudinais.

ABSTRACT

The abundance of classroom exams and its close relation with the actual curriculum justify an investigation of its use in research. The aim of the present one is to investigate some possibilities and limitations of common classroom assessment to track students' growth of competence in the subject of physics. Some characteristics of classroom assessments are discussed, its relation with competence and some possible advantages of its employ in contrast with other assessments or systemic evaluations. In the analysis, two kinds of assessment are used: trimestral grades and dichotomous items. For each, a longitudinal multilevel analysis is carried out and the results are compared and measured up to the literature. These comparisons indicate that classroom assessments can be used in studies that track growth of competence. There are three reservations, however, against the use of these practices: it can lead to a vague conception of competence; it can present difficulties for equalization; and it might not be adequate to a unidimensional approach. All of these reservations are sprung by limitations of the methodological design imposed by the ethics of the educational practice. Finally, study limitations are discussed.

Keywords: Competence in physics; classroom assessments; school grades; longitudinal studies

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Esta pesquisa tem como objetivo investigar algumas possibilidades e limitações do uso de avaliações escolares ordinárias para estudar a evolução de competência em física.

Pesquisas que investigam evolução da competência em diversos domínios usam, normalmente, testes padronizados (MULLER *et al.*, 2001; MA, WILKINS, 2002; POMPLUN, 2009). No entanto, o uso de testes para estudos longitudinais apresenta algumas dificuldades.

A primeira dificuldade se refere à possibilidade de coletar várias ondas de dados. Sabe-se que há problemas éticos em usar o tempo de aula para coletar dados. Isso inviabiliza a coleta de várias ondas de dados (necessária para um estudo longitudinal) em um curto espaço de tempo.

A segunda dificuldade está relacionada à possível falta de alinhamento entre os conteúdos dos testes com o currículo. Esse desalinhamento pode dificultar a detecção da evolução da competência no domínio de conhecimento em que se estuda o desenvolvimento.

Uma terceira dificuldade diz respeito ao engajamento dos estudantes nesses testes. Como as normas legais sobre ética na pesquisa com seres humanos no Brasil determinam que a participação em tais pesquisas deve ser voluntária, não se pode recompensar os estudantes de nenhuma forma (nem financeiramente, nem com distribuição de pontos, por exemplo). Mesmo nos Estados Unidos, onde as regras éticas para a concessão de incentivos à participação dos estudantes nas pesquisas são muito mais liberais que no Brasil, há relatos (ROESER, *et al.*, 2002) sobre a falta de engajamento dos alunos na realização desses testes. Roeser relata exemplos extremos de desengajamento deste tipo.

Todas essas dificuldades apontadas justificam a investigação de outras possibilidades para se estudar a evolução da competência. Uma dessas possibilidades pode ser o uso das avaliações escolares ordinárias, em suas diversas formas, com suas vantagens e limitações.

As avaliações escolares são, muitas vezes, feitas a cada mês, a cada quinzena ou até mesmo a cada semana. Assim, o uso dessas avaliações como dados de pesquisa pode possibilitar investigações de evolução em períodos de um ano ou menos, o que seria difícil (ou até mesmo impossível) por meio de testes padronizados ou avaliações feitas pelos sistemas de ensino.

Desde que o professor não subordine as suas decisões sobre o planejamento ou condução do ensino às necessidades ou conveniência da pesquisa, o uso das avaliações ordinárias pode vir a minimizar o problema do desengajamento, já que as avaliações fazem parte da disciplina e são usadas para se decidir sobre a aprovação do estudante.

Dessa forma, o presente estudo será guiado pela seguinte questão: *é possível usar notas escolares para estudar a evolução da competência em física?*

No próximo capítulo serão discutidos os referenciais teóricos e metodológicos usados na investigação. No terceiro capítulo, serão descritos o contexto escolar e a metodologia da pesquisa. No quarto capítulo, serão apresentados a análise dos dados e os resultados obtidos. No quinto capítulo, serão discutidos os resultados, a questão da pesquisa será retomada e serão apresentadas algumas limitações da pesquisa.

CAPÍTULO 1: REFERENCIAIS TEÒRICOS

Neste capítulo serão apresentados os referenciais teóricos e metodológicos usados para abordar o problema. Começarei com uma discussão acerca do que seria uma “competência em física”. Então, passarei a uma discussão das práticas escolares de avaliação e das competências relacionadas a elas. Em seguida, discutirei as características de estudos longitudinais de mudança. Para finalizar, discutirei o modelo Rasch como um método para inferir medidas de competência a partir do desempenho observado. Todas essas discussões serão breves e um entendimento mais profundo de cada um desses assuntos pode ser obtido através das referências bibliográficas citadas em cada seção.

1.1. COMPETÊNCIA EM FÍSICA

Este estudo tem como objetivo a investigação de algumas possibilidades do uso de avaliações escolares ordinárias para estudar a evolução da competência em física. O ponto de partida deve ser, então, a conceituação do que é “competência em física”.

Koeppen e seus colaboradores (2008) definem “competências” como “disposições contexto-específicas que são adquiridas e necessárias para lidar com sucesso com situações ou tarefas específicas em um domínio” (KOEPPEN, *et al*., 2008). A competência, portanto, diz respeito a um domínio específico. Além disso, um sujeito pode aumentar sua competência no domínio, ao aprender conhecimentos e habilidades específicos daquele domínio (como ocorre na escola).

As competências, ainda segundo Koeppen e seus colaboradores (2008), se diferenciam das habilidades cognitivas, que são traços quase independentes do domínio e difíceis de ser modificadas por meio de instrução.

Pode-se, então, em princípio, pensar em competência em física como a "capacidade de resolver questões e problemas de física". Fica subentendido que física, no contexto desta dissertação, não se refere a toda a ampla área de atuação dos físicos profissionais, mas apenas ao domínio mais restrito da física escolar. Assim, questões e problemas de física são aqueles e aqueles que podem ser atacados de forma produtiva com os conhecimentos e habilidades aprendidos e ensinados nas disciplinas de física no nível do ensino médio.

Weinert (1999, 2000, apud KOEPPEN, *et al*, 2008) apresenta vários argumentos para que o termo competência seja restrito a aspectos cognitivos e propõe que aspectos motivacionais ou afetivos devam ser avaliados como construtos separados.

Apesar de concordar até certo ponto com essa proposição, penso que tal separação pode ser inviável em algumas situações. Como exemplo, as pesquisas envolvendo alguns tipos de avaliações escolares, nas quais o dado observado (a "nota") já é, por natureza, uma mistura de aspectos cognitivos, motivacionais e afetivos (McMILLAN, 2001, 2003). Por outro lado, a conceituação de Koeppen e seus colaboradores (2008) não se limita às disposições cognitivas.

Opto, portanto, por trabalhar com uma conceituação de competência que pode envolver todos esses aspectos. Essa conceituação, um pouco vaga por um lado, permite que o termo adapte seu significado ao contexto. Assim, adoto a conceituação de Koeppen e seus colaboradores (2008), entendendo que a competência em um domínio pode ser vista como um *conjunto de estados e traços latentes específicos do domínio*.

Assumo, ainda, que seja possível medir a competência em um domínio de um sujeito usando modelos psicométricos, a partir do desempenho observado em tarefas ou testes específicos do domínio.

É claro que diferentes tipos de tarefas podem exigir competências diferentes, mesmo dentro de um mesmo domínio. Em última instância, a mínima modificação de um contexto pode levar à exigência de outras habilidades ou conhecimentos e, portanto, de outra competência. Pode-se então perguntar: até onde se deve considerar duas tarefas como fazendo parte do mesmo domínio? Essa é uma pergunta para a qual não há uma resposta única. Não se pode definir domínio a partir da conceituação de competência, uma vez que esta foi definida usando, justamente, a definição de domínio. Entendo que não há limites definidos para o que é ou não parte de um mesmo domínio. Tais limites dependem do propósito que se tem para o uso do termo. Por exemplo, pode-se considerar “ciências” como um domínio; pode-se considerar “física” como um domínio; ou pode-se considerar “mecânica” como um domínio. Quanto mais estreito o domínio, mais precisão e menos abrangência se tem.

Nesta pesquisa, usam-se dois indicadores da competência em física: escore em provas de itens dicotômicos (do tipo “verdadeiro ou falso”) e notas trimestrais. Este pesquisador está ciente de que essas duas “tarefas” não exigem exatamente a mesma competência: a competência exigida para se ter uma nota trimestral de física elevada envolve maior número de fatores de natureza motivacional e afetiva do que a competência exigida em provas de itens dicotômicos. Poder-se ia questionar o tratamento dessa competência para as notas trimestrais como “competência em física”. No entanto, entendo que essa competência é, sim, uma espécie de competência em física, pois envolve vários aspectos relativos a esse domínio, tais como conhecimentos e habilidades para responder questões fechadas de física, conhecimentos e habilidades para responder questões abertas de física, engajamento, interesse e persistência nas atividades de física, participação nas aulas de física. Assim, mesmo os aspectos não-

cognitivos dessa competência estão, em alguma medida, relacionados ao domínio “física”. Uso, portanto, o termo “competência em física” em ambos os casos (notas trimestrais e provas fechadas), ciente de que o significado muda (mas não de forma excessiva) de um caso para o outro.

1.2. AVALIAÇÕES ESCOLARES

A atribuição, pelo professor, de uma “nota”, ou conceito aos alunos é uma prática tradicional e antiga. Os critérios usados para se atribuir uma “nota trimestral” englobam vários aspectos, que variam conforme o contexto. Alguns estudos encontraram diferenças de critérios entre professores de diferentes disciplinas e de diferentes níveis de ensino (McMILLAN, 2001, 2003). Mesmo entre professores da mesma disciplina e que lecionam para o mesmo nível de ensino, há uma grande variedade de critérios (e do peso dado a cada critério) para se atribuir uma nota.

No entanto, mesmo com toda essa variabilidade, verifica-se que pelo menos dois grupos de critérios são amplamente usados (McMILLAN, 2001). O primeiro está relacionado com a aprendizagem, em seu sentido mais tradicional. O segundo está relacionado a uma espécie de “merecimento” (esforço do aluno, sua freqüência às aulas, sua participação nas aulas, responsabilidade etc). Alguns tipos de avaliação refletem mais o primeiro grupo de critérios (provas de conhecimento, por exemplo), outros refletem o segundo (como notas de conceito e participação) e outras estão no meio do caminho (atividades ou exercícios que envolvem o uso de conhecimentos e habilidades, mas que têm grandes chances de serem realizadas desde que haja certo grau de engajamento).

É sabido que fatores emotivos e motivacionais que influenciam no engajamento afetam também o resultado de testes de desempenho (ROESER *et al*, 2002; SHAVELSON *et al*, 2002; BYRNES, MILLER, 2007; LAWRENZ *et al*, 2009). No entanto, estudos (SHAVELSON *et al*, 2002) mostram que o nível de engajamento está mais relacionado às "notas finais" do que ao desempenho em testes.

Estou ciente de que pode haver críticas ao uso de notas escolares como uma medida de competência, exatamente por envolver também aspectos que não se relacionam, em princípio, com a aprendizagem de conteúdos (JUSSIM, 1991; WENTZEL, 1991; McMILLAN, 2001, 2003; BROOKHART, 2003). Mas há, pelo menos, três argumentos em defesa do uso das "notas" como um construto válido para analisar competência.

O primeiro argumento baseia-se na conceituação de competência usada na presente pesquisa. Entendo que competência não significa apenas habilidades e conhecimentos cognitivos, mas um conjunto maior de habilidades e conhecimentos (em uso), incluindo aspectos cognitivos, emotivos e motivacionais. Dessa maneira, por refletir esse conjunto de atributos de forma mais completa, a nota pode ser um bom indicador da competência do aluno.

O segundo argumento é que a nota escolar está intimamente relacionada aos objetivos curriculares, aos objetivos do professor e à percepção dos alunos do que deles se espera (incluindo aspectos do conteúdo e de comportamento). Não aos objetivos declarados em documentos, mas aqueles de fato enfatizados pelo professor em sua ação cotidiana na sala de aula. Ao contrário do que ocorre em testes padronizados (avaliações sistêmicas ou testes para pesquisas), o conteúdo cobrado em avaliações escolares ordinárias reflete o que foi trabalhado em sala pelo professor e o que foi estudado pelos alunos (BROOKHART, 2003).

Um terceiro argumento fundamenta-se na aceitabilidade social da nota como indicador de aprendizagem. A nota é o critério usado pelos alunos para avaliar sua aprendizagem,

gerando, inclusive, comparações entre os próprios estudantes e criação de rótulos de quem é ou não um “bom aluno”. Essa aceitabilidade faz com que entre os objetivos escolares dos estudantes esteja uma “busca pela nota” (de alguma forma e em algum grau). Portanto, parece-me adequado que se avalie a competência a partir de algo que está dentre os objetivos dos alunos e que eles se esforçam (em algum grau) para obter. O mesmo nível de engajamento pode não ocorrer com os testes que pesquisadores aplicam em sala de aula.

Continuando o terceiro argumento, deve-se ter em vista que a nota escolar também é a referência que os pais usam para avaliar o progresso de seus filhos na escola, o que reflete uma aceitabilidade dela parte dos pais. E, principalmente, ela é o critério usado para decidir se o aluno está apto a progredir (“passar de ano”), o que evidencia a aceitabilidade da comunidade escolar com um todo (diretores, professores, alunos, pais, e outros).

1.3. ESTUDOS LONGITUDINAIS

O estudo da evolução da competência em física é, por natureza, um estudo longitudinal de mudança. Nesta seção discutirei as características que devem ter os estudos longitudinais de mudança e apresentarei o modelo estatístico que será usado na análise dos dados.

1.3.1. Sobre estudos longitudinais de mudança

Segundo Singer e Willett (2003), apesar do interesse em se estudar mudança através do tempo ser antigo, os métodos estatísticos para que se possa fazer isso de forma apropriada só se desenvolveram a partir da década de 80. Esses métodos são nomeados sob diversos rótulos: modelos multinível, modelos hierárquicos lineares, modelos mistos, modelos de

crescimento individual, modelos com coeficientes randômicos. Uma exigência fundamental para que se possa estudar bem a mudança de variáveis no tempo é ter dados longitudinais (dados coletados em diferentes ocasiões para os mesmo indivíduos) (SINGER e WILLETT, 2003).

De um ponto de vista estatístico, todas as pesquisas sobre mudança têm como núcleo o seguinte par de questões: (i) como a variável de interesse muda com o tempo; (ii) que fatores ajudam a explicar como essa mudança varia entre os indivíduos (SINGER e WILLETT, 2003). Cada uma dessas questões deve ser tratada com um modelo. A primeira deve ser tratada com o um modelo de regressão da variável de interesse (dependente) no tempo (modelo de nível 1). A segunda questão deve ser tratada com um modelo de regressão dos coeficientes do modelo de nível 1 em função de fatores relacionados aos indivíduos (modelo de nível 2). De acordo com Singer e Willett (2003), a meta de uma análise nível 2 é detectar heterogeneidade na mudança entre indivíduos e determinar a relação entre os preditores e a forma de cada trajetória individual de crescimento. Os dois modelos (de nível 1 e de nível 2) devem ser considerados conjuntamente e é esse conjunto que é chamado de "modelo multinível" para mudança (SINGER e WILLETT, 2003).

No entanto, nem todo estudo longitudinal é apropriado para uma análise da mudança. Para que se faça um estudo de mudança, a pesquisa deve ter três características importantes (SINGER E WILLET, 2003):

i - Três ou mais ondas de dados.

> Por décadas, os pesquisadores acreditaram erroneamente que estudos com duas ondas eram suficientes para estudar mudança, porque eles conceituavam "mudança" de forma estreita, como "incremento": a simples diferença entre escores medidos em duas ocasiões de medida. (SINGER e WILLETT, 2003, p.10)

Segundo Singer e Willett (2003) o incremento não pode descrever o processo de mudança. Primeiro, porque a simples diferença de dois escores não contém nenhuma informação sobre a forma da mudança. Segundo, porque com apenas duas ondas de dados não se pode distinguir uma mudança real de um simples erro de medida. “Em termos estatísticos, estudos com duas ondas não podem descrever trajetórias individuais de mudança e confundem mudança verdadeira com erro de medida” (SINGER e WILLETT, 2003).

Quanto mais ondas de dados se coletam, mais informações se obtêm sobre o processo de mudança, sendo três o número mínimo. Com três ondas, temos de nos restringir a analisar trajetórias como se fossem lineares, mas podemos avaliar a qualidade do ajuste, ou seja, podemos estimar qual percentual da variância pode ser explicado pelo modelo intra-individual.

ii - Uma métrica sensível para o tempo:

A variável temporal deve ser medida em uma escala apropriada. A escolha adequada depende do contexto da pesquisa. Para algumas pesquisas, a “idade” pode ser uma boa escolha para a variável temporal. Em outros casos, a “série” pode ser mais adequada. Segundo Singer e Willett (2003), devemos escolher uma métrica para o tempo que reflita o ritmo esperado da mudança da variável dependente, com a única restrição de que, assim como o próprio tempo, a variável temporal seja estritamente crescente (nunca diminua com o tempo). Além disso, uma escolha do ponto inicial (ponto zero) pode proporcionar uma interpretação mais clara e mais direta dos resultados. No presente caso, por exemplo, uso o “trimestre” como variável temporal e o início do ano letivo como o ponto inicial.

É importante que o espaçamento entre as ocasiões de medida não seja pequeno demais, caso em que não seria apto a captar qualquer mudança substancial, nem grande demais, ao ponto de não captar nenhum detalhe do processo.

iii - Uma variável dependente contínua que muda sistematicamente através do tempo:

A variável dependente deve ter características de uma medida intervalar, ou seja, diferenças entre pares de valores, com o mesmo espaçamento na escala, devem ter o mesmo significado (ver a seção “Modelo Rasch de medida” mais adiante). Além disso, “a escala, a validade e a precisão da variável dependente devem ser mantidas através do tempo” (SINGER e WILLETT, 2003, p.13).

A escala ser mantida no tempo significa que um valor para a variável em uma ocasião tem o mesmo significado que o mesmo valor em outra ocasião. Isso pode ser conseguido com um método de equalização ou de calibração adequado.

1.3.2. O modelo multinível

Assumindo que todas essas exigências estão atendidas, podemos então escrever o modelo multinível. O modelo é chamado multinível porque pode ser separado em duas partes diferentes: uma para analisar mudanças do indivíduo no tempo (nível 1) e outra para analisar variação da mudança entre os indivíduos. O modelo de nível 1 é um modelo de regressão da variável de saída em função da variável temporal. O modelo de nível 2 é um conjunto de equações, no qual os coeficientes do modelo de nível 1 assumem o papel de variáveis dependentes e fatores relacionados aos indivíduos assumem o papel de variáveis independentes.

Um exemplo ilustrativo:

No caso de uma dependência linear com o tempo, podemos ter, para o modelo de nível 1:

$$Y_{ij} = [B_{0j} + B_{1j} \times TEMPO_i] + e_{ij}$$

Onde Y_{ij} é a variável dependente medida para o sujeito j, na ocasião i, a expressão entre colchetes, B_{0j} + B_{1j} x $TEMPO_i$, é a "trajetória verdadeira" prevista para o sujeito j. O coeficiente B_{0j} é o valor de Y_{ij} quando a variável "$TEMPO_i$" é nula, ou seja, é o intercepto da trajetória verdadeira. O coeficiente B_{1j} é a inclinação dessa trajetória (que mede o incremento na variável dependente para um acréscimo de uma unidade da variável "TEMPO"). O coeficiente e_{ij} é o resíduo (diferença entre o valor verdadeiro e o valor observado para o sujeito j na ocasião i). Esse resíduo pode ser interpretado como um erro inerente ao processo de medida ou uma variação da variável dependente não explicada pelo modelo. O modelo assume que o conjunto de resíduos (para todas as ocasiões e todos os sujeitos) tem uma distribuição normal com média zero e variância σ_e^2.

Suponha que queiramos investigar se uma certa variável "FEMININO" (que assume o valor "1" se o sujeito é do sexo feminino e "0" se é do sexo masculino) influencia no intercepto ou na inclinação prevista. Então, podemos escrever o modelo de nível 2:

$$B_{0j} = B_{00} + B_{01} \times FEMININO_j + u_{0j}$$
$$B_{1j} = B_{10} + B_{11} \times FEMININO_j + u_{1j}$$

Nesse modelo, a variável “FEMININO” aparece como preditora, tanto do intercepto (B_{0j}) quanto da inclinação (B_{1j}) da trajetória verdadeira dos indivíduos. O coeficiente u_{0j} é a diferença entre o intercepto (da trajetória verdadeira) do sujeito j e o intercepto médio de seu grupo (B_{00} para o grupo masculino e $B_{00} + B_{01}$ para o grupo feminino). O coeficiente u_{1j} é a diferença entre a inclinação da trajetória verdadeira do sujeito j e a inclinação média de seu grupo (B_{10} para o grupo de sujeitos do sexo masculino e $B_{10} + B_{11}$ para o grupo de sujeitos do sexo feminino).

O modelo assume que coeficientes u_{0j} e u_{1j} são ambos distribuídos normalmente, com média zero. As variâncias são respectivamente σ_0^2 e σ_1^2 e a covariância é σ_{01}.

Apesar da separação do modelo em dois conjuntos de equações (nível 1 e nível 2) facilitar a sua interpretação, a maioria dos softwares (inclusive o software que usamos: MLwiN) faz as estimativas utilizando um modelo composto. Para obtermos o modelo composto, basta substituir as equações de nível 2 no modelo de nível 1:

$$Y_{ij} = [B_{00} + B_{01} \times FEMININO_j + B_{10} \times TEMPO_i + B_{11} \times FEMININO_j \times TEMPO_i] + [u_{0j} + u_{1j} \times TEMPO_i + e_{ij}]$$

A parte da equação contida no primeiro colchetes é chamada de parte fixa do modelo. É importante notar que a influência da variável “FEMININO” na inclinação aparece como uma interação entre a variável de nível 1, “TEMPO”, e a variável de nível 2, “FEMININO”. A parte contida no segundo colchetes é chamada de parte randômica e para ela é suposta a estrutura de variância já mencionada.

Testando o ajuste de modelos:

A construção de um modelo multinível para a mudança nem sempre é um processo linear. Durante o processo, algumas variáveis são incluídas ou retiradas do modelo e este é, então, ajustado (utilizando-se um software) para verificar se as variáveis incluídas são bons preditores da mudança ou da variação interindividual nas trajetórias. Nesse processo, temos sempre que testar o ajuste dos modelos. Há algumas formas de verificar se a inclusão de novas variáveis melhorou o ajuste do modelo.

Uma delas é o uso da estatística "desviância" para modelos aninhados. Dizemos que dois modelos são "aninhados" se conseguimos transformar um modelo no outro apenas fazendo com que um ou mais coeficientes se iguale a zero. Nesse caso (e somente nesse caso) podemos comparar os dois modelos a partir da diferença nos valores da estatística desviância (SINGER e WILLETT, 2003). A diferença da desviância de dois modelos aninhados tem uma distribuição qui-quadrada, com a quantidade de graus de liberdade igual à diferença no número de parâmetros entre os dois modelos. Dessa forma, para testar se um modelo se ajusta melhor que o outro, basta fazer um teste qui-quadrado com o valor da diferença entre as desviâncias dos dois modelos (com o número de graus de liberdade dado pela diferença na quantidade de parâmetros). Se o teste fornecer um valor *p* menor que 0,05, consideramos que o ajuste foi significativo e o modelo com menor valor da desviância é o que melhor se ajusta (melhor explica a variância encontrada).

Outra forma de avaliar a melhoria trazida pela introdução de uma nova variável é a análise da variância dos coeficientes randômicos de nível 2. Como esses coeficientes correspondem à variância interindividual não explicada, uma diminuição nos seus valores representa uma melhor explicação da variância. Essa avaliação deve ser feita em conjunto

com o teste da desviância, e não como um primeiro critério para avaliar a qualidade da introdução de uma nova variável.

Outra forma de se avaliar se a variável introduzida traz informação relevante para a análise é verificar se o próprio coeficiente estimado é ou não significativo, mor meio de um simples teste *z* (dividir seu valor estimado pelo erro padrão).

1.4. MODELO RASCH DE MEDIDA

1.4.1. Medidas em Ciências Sociais

Uma das exigências para um estudo longitudinal de mudança, como já foi mencionado, é uma métrica para a variável dependente que seja estável no tempo. Quando falamos em uma escala que se mantém constante no tempo, podemos pensar em uma escala de medida de comprimento, por exemplo. Com uma régua graduada em milímetros, podemos medir o comprimento de vários objetos em diferentes momentos. Todos concordam que podemos comparar essas medidas sem maiores problemas, desde que a temperatura não varie muito, caso contrário, precisamos fazer as correções da escala para as variações de temperatura. Além disso, se constatamos que um objeto A é mais comprido que um objeto B, usando uma régua R, esperamos chegar à mesma conclusão usando outra régua R' e esperamos, também, chegar à mesma conclusão em qualquer instante de tempo, mantida constante a temperatura. Ninguém discute que essa é uma medida "unidimensional", e que seu resultado depende apenas de propriedades dos objetos que estão sendo medidos.

Nas ciências humanas, no entanto, o processo de medição é mais complicado. Primeiramente, porque estamos lidando com construtos teóricos não observáveis (como inteligência, habilidade, ou proficiência). Embora nas ciências exatas a maioria das medidas também seja de grandezas não diretamente observáveis (por exemplo: temperatura, pressão, força, corrente elétrica, diferença de potencial, entre muitas outras grandezas), as definições dessas grandezas e, por extensão, a teoria que conecta as observações às medidas têm uma clareza maior do que a dos construtos de áreas como a psicologia, por exemplo. Além disso, não é simples construir uma escala de valores com propriedades semelhantes às das escalas de comprimento ou de temperatura, dada a complexidade do ser humano.

Em meados do século XX, estudando a proficiência de jovens em leitura, o matemático dinamarquês Georg Rasch desenvolveu um modelo probabilístico que pode ser usado para se atribuir medidas (com propriedades semelhantes às medidas feitas com a régua) a qualidades psicológicas latentes, por meio do desempenho observado em testes. Dessa forma, o modelo usa dados observáveis (escore obtido nos itens de um teste) para inferir números para construtos teóricos não observáveis (proficiência do sujeito e dificuldade do item). Esse modelo tem como pressuposto que: (i) a proficiência de um sujeito não muda durante o teste; (ii) a resposta de um sujeito a um item não dependa da sua resposta a nenhum dos outros itens do teste; (iii) o teste seja unidimensional, quer dizer, a resposta a cada um dos itens do teste dependa apenas de uma única habilidade.

Quanto ao primeiro pressuposto, ele pode não ser verdadeiro, já que a pessoa pode aumentar sua proficiência (aprender alguma coisa) durante o teste. Ou seja, o teste pode ter (mesmo sem intenção) um caráter formativo.

O segundo pressuposto está relacionado ao primeiro: se a pessoa aprende (aumenta sua proficiência) ao responder um item do teste, isso pode influenciar na resposta aos outros itens.

Quanto ao terceiro pressuposto, de fato, nenhum teste pode atender plenamente ao requisito de unidimensionalidade. O resultado em um teste de lógica, por exemplo, não depende apenas do construto que se quer medir, mas de muitos outros fatores, entre eles, o próprio domínio do idioma em que o teste foi escrito. No entanto, em algumas populações, esses outros fatores podem não ter a variabilidade suficiente para serem captados pelo teste e, dessa forma, podemos considerar o teste como suficientemente unidimensional (para essa população) (RECKCASE, 2009). Nesse caso, o teste pode ser analisado com o modelo Rasch, gerando medidas em uma escala intervalar.

Mas, se pensarmos com mais cuidado, perceberemos que, mesmo no caso das ciências exatas, nenhum processo de medição é, em última instância, absolutamente unidimensional.

Pensemos no ato de medir o comprimento de um objeto com uma régua milimetrada como sendo um teste, ou melhor, um conjunto de testes. Por exemplo, se o limite do objeto está além da marca de “50mm”, podemos dizer que ele “passou” no teste: “ser ou não maior que 50mm”. Se o limite do objeto está além da marca de “51mm”, podemos dizer que ele “passou” no teste: “ser ou não maior que 51mm”. No entanto, a probabilidade de passar em cada um desses testes não depende apenas de propriedades intrínsecas do objeto. Os resultados desses testes dependem de vários outros fatores, como a temperatura que ele se encontra, a temperatura que a régua se encontra, a forma como o sujeito que faz a medição posiciona a régua, a precisão das marcas da régua, entre outros fatores. Em alguns casos, a própria interpretação subjetiva de ver o limite do objeto além da marca da régua pode não ser tão óbvia. Dessa forma, é possível que um objeto, em certo instante, passe no teste “ser maior que 50mm” e, em outro momento, não. É possível, até mesmo, que, em um momento, constate-se que “o objeto A é maior que o B”, e, em outro, não se alcance a mesma constatação. É claro que esses fatores influem muito pouco na medida do comprimento de um objeto e a régua não capta sensivelmente a variabilidade desses fatores. Contudo, nas ciências humanas e sociais, a influência de outros fatores na medida é muito maior, mas, qualitativamente, o processo de medida é análogo. Outra analogia com grandezas físicas (em alguns aspectos, melhor e mais completa) se encontra em BOND e FOX (2007, p.12).

Em suma, em qualquer caso real, os três pressupostos do modelo Rasch discutidos não serão atendidos completamente. No entanto, é possível haver situações em que elas sejam razoavelmente atendidas. Testes estatísticos podem ser usados para determinar se o afastamento dos pressupostos foi grande o suficiente para causar problemas nas medidas inferidas pelo modelo.

1.4.2. O Modelo Rasch para Itens Dicotômicos

O modelo Rasch considera que a probabilidade de acerto de um item dicotômico, construído para medir certa proficiência, depende apenas da dificuldade desse item e da proficiência do sujeito que o responde.

O modelo é

$$\ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = B_i - D_j$$

Onde B_i é a proficiência do sujeito i, D_j é a dificuldade do item j e P_{ij} é a probabilidade do sujeito i acertar o item j (1- P_{ij} é a probabilidade de erro).

Resolvendo a equação para *P*, obtemos (com uma notação mais completa):

$$P_{ij}(x_{ij} = 1/B_i, D_j) = \frac{e^{(B_i - D_j)}}{1 + e^{(B_i - D_j)}}$$

Onde P_{ij} ($x_{ij} = 1|B_i$, D_j) é a probabilidade da pessoa i obter escore x = 1 (ao invés de x = 0) no item j, dados a proficiência da pessoa B_i e a dificuldade do item D_j. Essa probabilidade é igual à base do logaritmo natural (*e* = 2,7183...) elevada à diferença entre *Bi* e *Dj* e depois dividida pelo mesmo valor somado à unidade. É importante notar que a probabilidade de uma pessoa i acertar ou não um item j, depende da diferença entre a proficiência *Bi* (considerada como a qualidade que está sendo medida pelos itens) da pessoa e a dificuldades *Dj* do item. Quanto maior essa diferença, maior a probabilidade de acerto (escore x = 1).

O gráfico abaixo representa a probabilidade de acerto de um item em função da diferença entre a proficiência do sujeito e a dificuldade do item.

Gráfico 1: Probabilidade de acerto de um item em função da diferença Bi – Dj

A estimativa das dificuldades dos itens e das proficiências dos sujeitos é feita por um processo numérico de iteração. Nesse processo, são feitos sucessivos ajustes nas estimativas das proficiências dos sujeitos e das dificuldades dos itens. Ao final do processo, a soma das probabilidades de acerto de cada item por um sujeito deve ser igual ao escore observado desse sujeito. Também a soma das probabilidades de acerto de cada sujeito, em determinado item, deve ser igual ao escore observado desse item (total de acertos nesse item). Isso para todos os sujeitos e itens. (Um algoritmo de iteração pode ser encontrado em MEAD (2008)).

É pertinente questionar sobre a possibilidade de se tratar a proficiência inferida pelo modelo como uma competência, dado que este termo (conforme foi conceituado) se refere a um construto, por natureza, multidimensional. De fato, ao assumir que poderei usar o modelo Rasch para estimar a competência em física, estarei também assumindo que essa competência pode ser tratada, aproximadamente, como unidimensional, o que pode parecer contraditório.

Para tentar esclarecer esse ponto, usarei uma analogia com a composição de um material como o granito: o granito é formado por diversos tipos de minerais e se o analisarmos com um microscópio, é possível que focalizemos partes mais concentradas de um ou outro mineral. Se compararmos essas pequenas partes, não poderemos dizer que se trata de um mesmo material. Mas, desde que olhemos para ele de uma perspectiva mais ampla, podemos tratar toda aquela mistura simplesmente como granito. Da mesma forma, assumirei que, olhando para a competência de uma perspectiva mais ampla, é possível estudá-la como um construto complexo, mas cuja complexidade, por não variar muito em sua composição, pode ser tratada aproximadamente como um conjunto único que pode ser medido, sem maiores problemas, em uma escala unidimensional.

A discussão desta seção justifica a escolha do modelo Rasch para inferir medidas para a competência em física dos sujeitos. Dessa forma, será obtida uma variável em uma escala intervalar e estável que, como foi discutido na seção anterior, é uma exigência fundamental para um estudo da mudança.

CAPÍTULO 2: METODOLOGIA

2.1. CONTEXTO

O estudo foi realizado com dados de uma Escola de Educação Básica e Técnica federal, situada em Belo Horizonte. Uma fração do alunado entra na escola por um concurso muito concorrido (geralmente mais de 30 candidatos por vaga), para fazer um curso técnico simultaneamente ao Ensino Médio. A outra parte ingressa na escola automaticamente, após concluir a nona série do Ensino Fundamental em outra escola pertencente à mesma instituição. Os alunos que ingressam desta última forma cursam apenas o Ensino Médio.

O currículo de física da escola é recursivo, em espiral. Na terceira série, no ano de 2008, as aulas de física eram estruturadas da seguinte forma: (i) os alunos liam um breve texto sobre a atividade; (ii) o professor discutia as dúvidas com a turma; (iii) os alunos respondiam, em pequenos grupos e com consulta, a questões discursivas sobre o assunto estudado, chamadas “tarefas”; (iv) o professor corrigia algumas dessas questões; (v) os alunos respondiam, individualmente, a um pequeno teste objetivo sobre o assunto da aula.

A nota trimestral era composta pelas notas das tarefas, dos testes, de avaliações intermediárias (abordando apenas o assunto estudado no trimestre), de avaliações trimestrais (abordando todo o conteúdo estudado no ano, até o momento) e de pontos de participação e conceito (que incluía a presença como um dos indicadores de participação).

2.2. SUJEITOS DA PESQUISA E OBTENÇÃO DOS DADOS

Serão analisados dados referentes a 147 alunos que cursaram a terceira série em 2008. Esses alunos se dividiam em seis turmas: três delas (turmas T1, T4 e T5) contendo alunos do curso de Instrumentação e do curso de Eletrônica; uma (turma T6) formada por alunos do curso de Química; uma (turma T3) por alunos do curso de Patologia Clínica; e uma (turma T2) formada por alunos que não ingressaram por concurso e não faziam (em sua maioria) nenhum curso técnico. As turmas de Instrumentação e Eletrônica eram turmas predominantemente masculinas, enquanto a turma de Patologia Clínica era uma turma predominantemente feminina.

Os dados analisados foram fornecidos pela secretaria da escola e pelo coordenador da disciplina. Esses dados são: notas trimestrais de todas as três séries do Ensino Médio, em física e em matemática, respostas dos alunos de três turmas (T4, T5 e T6) às três provas trimestrais de física (todas com itens do tipo “verdadeiro ou falso”), a turma a que cada aluno pertenceu na terceira série e o professor de física de cada turma da terceira série. Além disso, consegui dados da faixa de renda, escolarização do pai e escolarização da mãe de 112 desses alunos que responderam ao questionário sócio-econômico ao se inscreverem no vestibular da UFMG para o ano de 2009[1], doravante denominado Questionário Sócioecônomico do vestibular (QSEV).

[1] Os dados foram liberados pela Copeve atendendo ao pedido do Prof. Arnaldo Vaz, a quem agradeço.

2.3. CRIAÇÃO DAS VARIÁVEIS

2.3.1. Variável temporal

O tempo (ou outra variável temporal) é a principal variável independente em um estudo de mudança, sendo o primeiro candidato a explicar a variação da competência dos sujeitos entre diferentes ocasiões.

Foi criada uma variável temporal baseada na divisão do ano letivo em trimestres. Essa variável vale 1 para eventos que ocorreram no primeiro trimestre, 2 para eventos que ocorreram no segundo trimestre e 3 para eventos que ocorreram no terceiro trimestre. Ela foi denominada “TEMPO”.

2.3.2. Variáveis dependentes

Foram criadas duas variáveis dependentes com indicadoras da competência em física. Uma que chamei de “R(POMP)” e a outra que chamei de “R(ITENS)”.

2.3.2.1. Variável R(POMP):

Essa é a variável que pretende medir a competência dos estudantes a partir da nota trimestral obtida por eles.

Como o primeiro trimestre valia 30 pontos e os outros dois, 35 pontos, dividi a nota de cada trimestre pelo valor total, obtendo assim um número decimal que representa o percentual em relação ao máximo valor possível, que chamamos POMP (sigla para Percent Of Maximum Possible score). (COHEN *et al*, 1999)

Converti o POMP obtido em um conceito. Os POMPs inferiores a 0,60 foram classificados como conceito “D”; os que eram maiores ou iguais a 0,60 e menores que 0,70 foram classificados com “C”; os que eram maiores ou iguais a 0,70 e menores que 0,80 foram classificados como “B”; os que eram maiores ou iguais a 0,80 foram classificados como “A”. (inicialmente, utilizei seis conceitos – de “A” a “F” – mas, devido ao ínfimo número de alunos no mais alto e no mais baixo, mudei a categorização para a que foi apresentada).

Esses quatro conceitos foram tratados como se fossem escores obtidos em um “teste” com três “itens” dicotômicos. A “resposta” ao “item1” era considerada “correta” se o conceito fosse “A”, “B” ou “C” e “errada” se fosse “D”. A “resposta” ao “item2” era considerada “correta” se o conceito fosse “A” ou “B”, e “errada” se fosse “C” ou “D”. A “resposta” ao “item3” era considerada correta de o conceito fosse “A” e “errada” se fosse “B”, “C” ou “D”.

No entanto, devido a possíveis diferenças entre o grau de leniência de diferentes professores e também das possíveis diferenças na distribuição de pontos entre diferentes turmas, não podemos considerar que todos os sujeitos responderam aos mesmos “itens”.

Por isso, considerei que cada turma respondeu, em cada trimestre, a testes diferentes. Dessa forma, fiquei com 54 “itens” em 18 “testes” (um teste para cada um dos três trimestres, para cada das seis turmas). É claro que cada sujeito só tem escore em nove itens (3 testes): três itens (1 teste) em cada trimestre.

Para equalizar esses testes, procedi de acordo com as orientações de Linacre (2010) para uma “Equalização Virtual de Formas de Testes”. Segui os seguintes passos:

1) Primeiramente, verifiquei a possibilidade de equalizar os dezoito diferentes testes.

    a) Calculei as dificuldades dos itens, através de uma análise Rasch (com o software Winsteps) para cada “teste” separadamente.

b) Escolhi arbitrariamente o “teste” do primeiro trimestre da turma M-31 como o teste de referência “a”.

c) Escolhi itens similares entre cada um dos outros “testes” e o “teste” de referência (considerei similares os itens correspondentes aos mesmos conceitos).

d) Calculei as médias e os desvios padrão das dificuldades obtidas pela análise separada para cada “teste”. Como só havia três itens comuns para cada par de testes, o desvio padrão foi calculado pela equação $s = \frac{(x_2 - x_1) + (x_3 - x_2)}{\sqrt{3}}$ (onde $x_1 \leq x_2 \leq x_3$), que é o limite superior para o desvio padrão em uma amostra com n = 3 (JOARDER e LATIF, 2006).

e) Tracei, para cada “teste” j, a reta que passa pelos pontos (Ma, Mj) e (Ma+Sa, Mj+Sj). Sendo Ma a média das dificuldades dos itens comuns obtidas na análise do “teste” a (de referência); Mj a média das dificuldades para os itens comuns obtidas na análise do “teste” j; Sa o desvio padrão das dificuldades os itens comuns na análise do teste de referência; e Sj o desvio padrão das dificuldades dos itens comuns na análise do “teste” j. Segundo Linacre (2010), se a inclinação dessa reta (dada por Sb/Sa) for próxima da unidade, os testes podem ser equalizados.

f) Construí intervalos de confiança (95%) para as estimativas de Ma, Mj, Ma+Sa e Mj + Sj, de acordo com TRIOLA (2008).

g) Construí, com a ferramenta “desenho” do Excel, uma reta de inclinação 1.

h) Verifiquei, visualmente, se essa reta poderia ser disposta de forma a cruzar os intervalos de confiança criados para os pontos. Caso isso ocorra, não se pode afirmar que a inclinação da reta construída no passo “1.e” seja diferente de 1.

i) Em todos os casos, verifiquei que era possível haver uma reta de inclinação 1 que cortasse os intervalos de confiança construídos, indicando que não se poderia afirmar que a inclinação da reta que passa pelos pontos (Ma,Mj) e (Ma+Sa, Mj+Sj) fosse diferente da unidade. Portanto, os "testes" poderiam ser equalizados ao teste de referência.

2) A equalização consiste em rodar a análise Rasch de cada teste, no Winsteps, fixando: (i) a média das dificuldades no valor do intercepto da reta com o eixo x (construída no passo "1.e"); (ii) a unidade da escala como sendo o inverso da inclinação dessa reta.

a) Utilizando os intervalos de confiança (ver passo "1.f") e a reta de inclinação 1 (ver passo "1.g") verifiquei que, em nenhuma dos "testes", era possível afirmar que o intercepto da reta com eixo das abscissas é diferente de zero. Isso significa que não se pode afirmar que eles não estivessem na mesma escala.
b) Como os valores dos interceptos (com o eixo x) não eram significativamente diferentes de zero, optei por não utilizá-los para fixar as escalas, com receio de que isso pudesse introduzir mais um artefato metodológico, sem necessidade.

Essa verificação me deu segurança para simplesmente entrar com todos os "testes" de uma só vez, em uma grande matriz, para a análise (sem fixar médias e escalas diferentes para cada um).

No entanto, o Winsteps não é capaz de avaliar uma matriz com padrão de respostas completamente determinístico, como a que foi criada, a partir dos conceitos trimestrais dos sujeitos (ver matriz abaixo). O motivo é que, nesse padrão, sempre há pelo menos um item com escore nulo ou máximo e/ou um sujeito com escore nulo máximo.

Tomemos a hipotética matriz de respostas abaixo como uma mera ilustração.

	ITEM1	ITEM2	ITEM3	ITEM4
SUJ1	1	0	0	0
SUJ2	1	1	0	0
SUJ3	1	1	1	0
SUJ4	1	1	1	1

Figura 1: tabela ilustrativa de um padrão de respostas determinístico

Nela há um sujeito (SUJ4) com escore máximo e um item (ITEM1) com escore máximo. O programa não consegue, inicialmente, estimar uma medida para competência do sujeito 4. Isso porque a estimativa inicial das medidas de competência envolve o logaritmo natural do escore total (r) do sujeito dividido pela diferença entre o máximo escore possível (M) e o escore total (r). Se r =0, teremos ln(0), se r = M, teremos ln(r/0). Algo semelhante ocorre para a estimativa dos itens.

O algoritmo faz com que esse sujeito (ou item) com escore total seja eliminado provisoriamente da matriz. Nesse exemplo, ao eliminar o sujeito 4, ficaremos com dois itens sem possibilidade de estimativa (Item 1 com escore máximo, e Item 4 com escore nulo). Assim, seguindo o algoritmo, o software descartará, provisoriamente, esses dois itens. Sobrarão apenas as linhas 1, 2 e 3 e as colunas 2 e 3 da matriz. Podemos perceber que nessa nova matriz reduzida, então, os sujeitos 1 e 3 deverão ser excluídos, pois seus escores são, respectivamente, zero e máximo. O processo de exclusão continuaria até que toda a matriz fosse excluída, não importa o seu tamanho.

Por isso, para que o software pudesse fazer a análise da nossa matriz (que é semelhante à matriz ilustrativa apresentada ampliada) acrescentei um “sujeito virtual”, como sugerido por Linacre (2010). Esse sujeito tem um padrão de respostas escolhido de modo a

não deixar que nenhum item tenha escore total ou nulo (o mesmo processo foi feito quando analisei os testes individualmente).

Além disso, lembremos que a análise dos parâmetros da reta obtida com os itens comuns (já descrita) não levou à conclusão de que os "testes" estavam equalizados. Ela levou apenas à conclusão de que não se pode afirmar que não estivessem equalizados. Mas a introdução do "sujeito virtual" pode fornecer uma garantia mais forte dessa equalização. Isso porque esse sujeito tem respostas em todos os "itens" de todos os "testes", conectando as partes da matriz que estariam separadas.

Entrei então com a matriz, contendo todos os sujeitos e todos os "testes", no software Winsteps. Todas as células vazias da matriz correspondem a situações em que o sujeito não "respondeu ao item" (por exemplo, sujeitos da turma T1 e "itens" de um "teste" para a turma T2) foram tratados como dados faltantes. Além disso, cada sujeito foi codificado de três formas. Portanto, o software tratou cada sujeito como sendo três diferentes sujeitos (um para cada ocasião). Isso foi feito para que pudessem ser estimadas, para um mesmo sujeito, competências diferentes para ocasiões diferentes.

A figura abaixo, meramente ilustrativa, tem o objetivo de dar uma idéia de como foi montada a matriz de respostas. Cada linha corresponde a um sujeito em uma ocasião e cada coluna corresponde a um item. A letra "R" indica que há um escore (0 ou 1) para o item e o espaço vazio significa que não há um escore para o item..

TURMA	SUJEITO	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12
TURMA X	Suj1(Oc1)	R	R	R									
ocasião1	Suj2(Oc1)	R	R	R									
	Suj3(Oc1)	R	R	R									
TURMA Y	Suj4(Oc1)				R	R	R						
ocasião1	Suj5(Oc1)				R	R	R						
	Suj6(Oc1)				R	R	R						
TURMA X	.Suj1(Oc2)							R	R	R			
ocasião2	.Suj2(Oc2)							R	R	R			
	.Suj3(Oc2)							R	R	R			
TURMA Y	Suj4(Oc2)										R	R	R
ocasião2	Suj5(Oc2)										R	R	R
	Suj6(Oc2)										R	R	R
	Sujeito virtual	R	R	R	R	R	R	R	R	R	R	R	R

Figura 2: estrutura de matriz usada para obter o R(POMP)

Após todo esse processo, fiquei com uma medida de competência em uma escala intervalar, obtida por meio das notas trimestrais dos sujeitos. Essa medida, que chamei de R(POMP), é uma das variáveis dependentes que será usada na análise longitudinal.

2.3.2.2. Variável R(ITENS):

Essa variável foi criada utilizando o modelo Rasch para analisar as respostas às provas trimestrais.

Características das provas trimestrais

A Trimestral1 foi aplicada no final do primeiro trimestre e havia dois tipos de prova. Cada um continha 96 itens, sendo que 68 dessem itens eram comuns às duas provas. Todo conteúdo abordado pode ser considerado como conteúdos de “Mecânica”.

A Trimestral2 foi aplicada no final do segundo trimestre, abordando conteúdos de mecânica e de eletricidade. Havia dois tipos de prova, ambos contendo os mesmos itens (em ordem diferente). A Trimestral2 tinha itens comuns com a Trimestral1.

A Trimestral3 foi aplicada no final do terceiro trimestre, abordando conteúdos de mecânica, eletricidade e eletromagnetismo. Havia quatro tipos de prova e a maioria dos itens aparecia em apenas dois deles (ocorrendo várias combinações). Além disso, essa prova continha várias questões comuns com a Trimestral1 e com a Trimestral2.

Analisando a qualidade dos itens

Primeiramente, analisei a matriz de respostas de cada prova separadamente, obtendo a competência dos sujeitos e as dificuldades dos itens. Com isso, obtive, para cada item, a curva de probabilidade de acerto (prevista pelo modelo Rasch) em função da competência. Separei, então, os sujeitos em “grupos de desempenho” (definidos a partir de faixas de escore total na prova) e calculei a média do escore percentual (escore observado dividido pelo máximo escore possível) de cada grupo em cada item. Calculei também a média das “competências” dos sujeitos de cada grupo.

Para cada item plotei, no mesmo gráfico (ver exemplo abaixo), os pontos observados (y = percentual observado de acertos do item para o grupo - que é o “escore médio” do grupo no item; x = competência média do grupo) e a curva de probabilidade de acerto do item em função da competência. As barras de erro para o percentual de acertos observado (escore médio) de cada grupo foram calculadas usando o erro padrão para a estimativa de uma

proporção (TRIOLA, 2008, p.259) multiplicada pelo fator 1,4 - como sugerido por Goldstein (1995). As margens de erro para a curva de probabilidade de acerto do item (curvas pontilhadas) são simplesmente as curvas de probabilidade de acerto (em função da competência) calculadas usando-se os limites do intervalo de confiança para a estimativa da dificuldade do item. Esse intervalo de confiança é a soma (ou diferença) da dificuldade estimada do item e do produto do erro padrão (fornecido pelo software) da dificuldade do item pelo fator 1,4.

Por meio de uma inspeção visual, verifiquei se, para cada grupo de desempenho, em cada item, a barra de erro dos pontos observados "cruzava" os limites de estimativa da curva de probabilidade (ver figura abaixo). Se isso ocorria para todos os grupos de desempenho, considerei que o item estava "funcionando bem". Ou seja, a probabilidade de acerto prevista para dada competência (curva) estava compatível com a porcentagem média de acertos observada para grupos com esse valor médio de competência (pontos), para todas as faixas de competência.

EXEMPLO

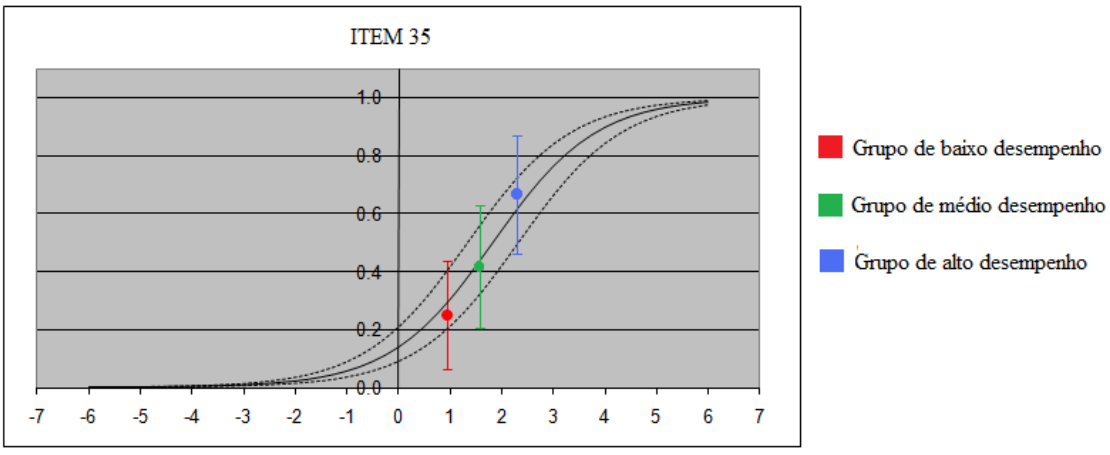

Figura 3: Exemplo de análise de ajuste de um item

Excluí os itens que não estavam adequados e re-analisei a prova, obtendo novas competências, novas dificuldades e novos grupos de desempenho. Verifiquei novamente a

adequação dos itens, seguindo os mesmos passos. Excluí, mais uma vez, os que não estavam bons. Todo esse processo foi feito para cada prova, separadamente, até que só restassem "bons itens". Portanto, após esse processo, para cada prova, todos os itens estavam avaliando a mesma competência que a prova como um todo. Ou seja, havia um bom indício de que as provas estavam avaliando algo que podia ser considerado razoavelmente unidimensional.

Equalizando as escalas

Para esse tipo de estrutura de testes, a sugestão de equalização feita por Linacre é a "Equalização de Itens Comuns". Guiado pelas orientações de Linacre (2010), segui os seguintes passos para verificar a possibilidade de equalização:

a) Calculei as dificuldades dos itens, por meio de uma análise Rasch (com o software Winsteps) para cada prova separadamente.
b) Escolhi, arbitrariamente, a prova "Trimestral3" como referência (por ter conteúdo mais abrangente).
c) Identifiquei os itens comuns entre cada um dos outras provas e a prova de referência.
d) Calculei as médias e os desvios padrão das dificuldades obtidas pela análise separada para cada prova
e) Tracei, para cada prova j, a reta que passa pelos pontos (Ma, Mj) e (Ma+Sa, Mj+Sj), onde Ma é a média das dificuldades dos itens comuns obtidas na análise da prova "a" de referência, Mj é a média das dificuldades para os itens comuns obtidas na análise da prova "j", Sa é o desvio padrão das dificuldades os itens comuns na análise da prova de referência e Sj é o desvio padrão das dificuldades dos itens comuns na análise da prova j. Segundo Linacre (2010), se a inclinação dessa reta (dada por Sb/Sa) for próxima da unidade, as provas podem ser equalizadas.

f) Construí intervalos de confiança (95%) para as estimativas de Ma, Mj, Ma+Sa e Mj + Sj, de acordo com TRIOLA (2008).

g) Construí, com a ferramenta “desenho” do Excel, uma reta de inclinação 1.

h) Verifiquei, visualmente, se essa reta poderia ser disposta de forma a cruzar os intervalos de confiança criados para os pontos. Caso isso ocorra, não se pode afirmar que a inclinação da reta construída no passo “1.e” seja diferente de 1.

i) Em todos os casos, verifiquei que era possível haver uma reta de inclinação 1 que cortasse os intervalos de confiança construídos, indicando que não se poderia afirmar que a inclinação da reta que passa pelos pontos (Ma,Mj) e (Ma+Sa, Mj+Sj) fosse diferente da unidade. Portanto, os testes poderiam ser equalizados ao teste de referência com o método proposto por Linacre (2010).

Por esse método, insere-se uma matriz completa, contendo as respostas de todas as provas, no software, para a análise. Os sujeitos são tratados separadamente, por ocasião (ver figura abaixo) e os itens comuns a duas ou mais provas são dispostos na mesma coluna.

A ilustração abaixo representa simplificadamente essa estrutura (nesse exemplo, há apenas duas ocasiões e os itens 1, 3 e 5 são comuns às duas):

	I1	I2	I3	I4	I5	I6	I7	I8	I9
Suj 1(Oc 1)	R	R	R	R	R	R			
Suj 2(Oc 1)	R	R	R	R	R	R			
Suj 3(Oc 1)	R	R	R	R	R	R			
Suj 1(Oc 2)	R		R		R		R	R	R
Suj 2(Oc 2)	R		R		R		R	R	R
Suj 3(Oc 2)	R		R		R		R	R	R

Figura 4: estrutura da matriz usada para obter o R(ITENS)

Montei então uma grande matriz (com estrutura semelhante à da figura), contendo as respostas de cada indivíduo (em cada ocasião) aos itens a que foi submetido. Nessa matriz, as linhas representavam um par sujeito-ocasião e as colunas representavam os itens. Antes de entrar com a matriz no software, fiz o mesmo tipo de análise de ajuste de itens relatado acima, para excluir os itens que não tiveram o mesmo funcionamento para todas as ocasiões. Excluí, ainda, itens que não eram comuns aos dois tipos da Trimestral1, para que ficasse com uma matriz menos esparsa (o que melhora as estimativas do software).

Após isso, inseri a matriz no software Winsteps e obtive uma medida de competência, que chamei de "R(ITENS)", para cada sujeito em cada ocasião.

Estimando competências para as turmas T1, T2 e T3

No entanto, como já mencionei, eu não tinha as respostas dos alunos das turmas T1, T2 e T3 para a Trimestral1. Portanto, a análise Rasch que descrevi não me forneceu medidas para a competência desses alunos na primeira ocasião. No entanto, eu tinha o escore total desses alunos nessa prova. Para obter estimativas para essas medidas, adotei o seguinte procedimento:

1) A partir do escore total e das dificuldades dos itens da Trimestral1 (obtidas a partir de uma análise Rasch somente dessa prova), usei o método de iteração proposto por MEAD (2008, p.28) para obter uma "medida provisória" de competência para os sujeitos.
2) Usei essa "medida provisória" para estimar um escore apenas nos itens da Trimestral1, que foram usados na análise da matriz completa. Chamei esse escore de "escore reduzido" (neste passo, ainda utilizei as dificuldades obtidas na análise da Trimestral1 feita separadamente).

3) Esse “escore reduzido” previsto foi arredondado para o inteiro mais próximo.
4) Obtive a medida de competência (estimada pela análise completa) para cada valor de escore, para os sujeitos da primeira ocasião (das turmas T4, T5 e T6). Como esses sujeitos da primeira ocasião responderam apenas aos itens da Trimestral1, esse escore corresponde ao “escore reduzido” – que foi definido como o “escore nos itens da Trimestral1 que entraram na análise completa”. Portanto, fiquei com a medida de competência para cada valor de escore.
5) A medida para dois sujeitos com mesmo o escore total e que responderam aos mesmos itens devem ser iguais. Portanto, pude estimar medidas de competência para alunos das turmas T1, T2 e T3. Fiz isso usando o valor de seus escores reduzidos (estimados nos passos “a”, “b”, “c”) e a relação entre o escore reduzido e a competência (obtida no passo “d”).

Dessa forma, fiquei com as medidas de competência em física, R(ITENS), obtidas a partir de provas com itens comuns, para todos os sujeitos e para todas as ocasiões.

2.3.3. Outras variáveis independentes (preditores)

As variáveis independentes mencionadas abaixo foram criadas para tentar explicar a variação existente nas variáveis dependentes.

Desempenho prévio em física

> “*Se eu tivesse que reduzir toda a psicologia educacional para apenas um princípio, eu diria isto: o fator singular mais importante que influencia o aprendizado é o que o aprendiz já sabe” (AUSUBEL, 1978, p.vi).*

Essa frase do psicólogo David Ausubel ilustra bem a importância que o conhecimento prévio tem na aprendizagem. Por isso, procurei obter um indicador relacionado ao conhecimento prévio.

Transformei as notas trimestrais de física da primeira e da segunda série em POMPs (percentuais do máximo valor possível). Para cada sujeito, somei os POMPs obtidos e dividi por seis (já que eram seis notas trimestrais). O valor obtido foi chamado de "desempenho prévio em física". Criei então uma variável categórica, com três categorias, relacionada a esse "desempenho prévio".

DPF: *é classificado como "baixo" se o desempenho prévio em física é menor que 0,70; é classificado como "médio" se o desempenho prévio em física é maior ou igual a 0,70 e menor que 0,85; é classificado como "alto" se o desempenho prévio em física é maior que 0,85.*

Esse não é um indicador apenas do conhecimento prévio, mas da competência em física (de uma forma mais geral) com que cada estudante chega à terceira série, além de incluir outros fatores, relacionados à maneira como as notas foram atribuídas, em física, nas séries anteriores.

Desempenho prévio em matemática

Como a matemática é uma ferramenta fundamental para a física, o nível de competência em matemática é um bom candidato a influenciar a aprendizagem de física. Por isso, optei por incluir um indicador da competência em matemática com que os alunos chegam à terceira série.

Transformei as notas trimestrais de matemática da primeira e da segunda série em POMPs (percentuais do máximo valor possível). Para cada sujeito, somei os POMPs obtidos e dividi por seis (já que eram seis notas trimestrais). O valor obtido foi chamado de "desempenho prévio em matemática". Criei então uma variável categórica, com três categorias, relacionada a esse "desempenho prévio".

DPM: *é classificado como "baixo" se o desempenho prévio em matemática é menor que 0,70; é classificado como "médio" se o desempenho prévio em matemática é maior ou igual a 0,70 e menor que 0,85; é classificado como "alto" se o desempenho prévio em matemática é maior que 0,85.*

Assim como no caso da física, esse é um indicador da competência (nesse caso em matemática) em com que cada estudante chega à terceira série, mas inclui também outros fatores relacionados à maneira como as notas foram atribuídas, em matemática, nas séries anteriores.

Gênero

Incluí uma variável para o gênero porque várias pesquisas já encontraram diferenças no desempenho em Ciências entre meninos e meninas (MULLER et al, 2001; BYRNES e MILLER, 2007; LAWRENZ et al, 2009, GRIGG et al, 2006) . Eis como foi criada a variável "GÊNERO":

GÊNERO: *assume o valor "0" para sujeitos do sexo masculino e "1" para sujeitos do sexo feminino.*

Escolaridade dos pais

Incluí a escolarização dos pais como um possível preditor, pois sua influência na aprendizagem de ciências já foi identificada em alguns estudos (CATSAMBIS, 1998, JOHNSON, 2009). Além disso, estudos apontam o nível sócio-econômico como preditor de desempenho (WHITE, 1992; MA e WILKINS, 2002; SIRLIN, 2003; BYRNES e MILLER, 2007) e, muitas vezes, a escolarização dos pais é um dos critérios usados para definir o nível sócio-econômico.

***Escolarização do pai*:**

A escolaridade do pai dos estudantes foi obtida por dados do QSEV, para 112 estudantes. O nível de escolaridade do pai foi declarado pelo estudante ao responder ao QSEV. Inicialmente, foram criadas quatro categorias: 0 – Fundamental incompleto; 1- Fundamental completo; 2- Médio completo; 3-Superior completo. Mas, como havia muito poucas pessoas nas categorias 0 e 1, juntei essas categorias à categoria 2.

Fiquei então com apenas duas categorias, que transformei em uma variável dicotômica:

E.Pai: *assume o valor "1" se o aluno declarou que seu pai completou o Ensino Superior e "0" se ele declarou que seu pai não completou o Ensino Superior.*

***Escolaridade da mãe**:*

A escolaridade da mãe dos estudantes foi obtida por dados do QSEV, para 112 estudantes. O nível de escolaridade da mãe foi declarado pelo estudante ao responder ao QSEV. Inicialmente, foram criadas quatro categorias: 0 – Fundamental incompleto; 1- Fundamental completo; 2- Médio completo; 3-Superior completo. Mas, como havia muito poucas pessoas nas categorias 0 e 1, juntei essas categorias à categoria 2.

Fiquei então com apenas duas categorias, que transformamos em uma variável dicotômica:

E.Mae: *assume o valor "1" se o aluno declarou que sua mãe completou o Ensino Superior e "0" se ele declarou que sua mãe não completou o Ensino Superior.*

Renda Familiar

A inclusão de uma variável para a faixa de renda também se justifica pelo resultado de pesquisas empíricas (WHITE, 1992; MA e WILKINS, 2002; SIRLIN, 2003; BYRNES e MILLER, 2007).

A renda familiar foi declarada no QSEV. O QSEV previa sete faixas de renda familiar: de um a dois salários mínimos, de dois a cinco salários mínimos, de cinco a dez salários mínimos, de dez a quinze salários mínimos, de quinze a vinte salários mínimos, de vinte a quarenta salários mínimos, mais de quarenta salários mínimos.

No entanto, na amostra utilizada, algumas dessas categorias estavam vazias (ou quase vazias) e optou-se por colapsar muitas delas. Sobraram então, apenas duas categorias que transformei em uma variável dicotômica:

RENDA: *assume o valor "0" se o aluno declarou renda familiar inferior a cinco salários mínimos e "1" se declarou renda superior a 5 salários mínimos.*

Professor

Em 2008, três professores lecionavam a disciplina "física" para as turmas da terceira série da escola pesquisada. Um deles, o "professor 1", era um professor efetivo da

universidade, já experiente. Os outros dois, “professor2” e “professor3”, eram jovens professores substitutos, com pouca experiência docente, sendo um deles este pesquisador.

Três categorias relacionadas aos professores foram criadas:

PROF1: *assume o valor “1” se o professor de física do sujeito era o “professor 1”, e “0” caso contrário.*

PROF2: *assume o valor “1” se o professor de física do sujeito era o “professor 2”, e “0” caso contrário.*

PROF3: *assume o valor “1” se o professor de física do sujeito era o “professor 3”, e “0” caso contrário.*

Turma

É plausível pensar que o efeito dos pares pode influenciar na aprendizagem escolar. Na verdade, há resultado de pesquisas que mostram esta influencia. Além disso, as turmas eram dividas de acordo com o curso técnico feito pelos alunos, podendo, portanto, refletir certo vocacionamento. A relação entre vocacionamento e aprendizagem de física já foi identificada em pesquisas (COELHO e BORGES, 2010).

Como já foi descrito, havia seis turmas de terceira série na escola, em 2008. Três dessas turmas, “T1”, “T4” e “T5” eram compostas de alunos que ingressaram na escola por meio de concurso e faziam cursos de Eletrônica ou Instrumentação. A turma “T3” era composta de alunos que ingressaram na escola por de concurso e faziam o curso de Patologia Clínica. A turma “T6” era composta de alunos que ingressaram na escola por meio de concurso e faziam o curso técnico de Química. A turma “T2” era composta por alunos que ingressaram automaticamente na escola, após a conclusão do Ensino Fundamental em uma escola pertencente à mesma Universidade. Esses alunos, em sua maioria, não faziam nenhum

curso técnico (com a exceção de poucos que entraram em vagas que surgiram, por desistência de outros).

Para identificar a turma à qual cada sujeito pertencia, foram criadas seis variáveis dicotômicas:

T1: *assume o valor "1" se o sujeito pertencia à turma "T1" e "0", em outros casos.*

T2: *assume o valor "1" se o sujeito pertencia à turma "T2" e "0", em outros casos.*

T3: *assume o valor "1" se o sujeito pertencia à turma "T3" e "0", em outros casos.*

T4: *assume o valor "1" se o sujeito pertencia à turma "T4" e "0", em outros casos.*

T5: *assume o valor "1" se o sujeito pertencia à turma "T5" e "0", em outros casos.*

T6: *assume o valor "1" se o sujeito pertencia à turma "T6" e "0", em outros casos.*

2.4. ANÁLISE DOS DADOS

A análise dos dados será dividida em quatro etapas: a primeira consistirá em uma análise exploratória; a segunda, na análise multinível a partir das notas trimestrais; a terceira será a análise multinível a partir do desempenho nas provas fechadas; a quarta será a comparação dos resultados obtidos nas duas análises multinível.

2.4.1. Análise exploratória

Na análise exploratória, serão examinadas as relações entre as variáveis R(POMP) e R(ITENS). Usando recursos gráficos, analisarei como as médias das medidas de competência

(R(POMP) e R(ITENS)) variam de ocasião para ocasião. Além disso, verificarei se há correlação entre elas.

2.4.2. Análise multinível utilizando as notas trimestrais

Utilizarei, então, um modelo de regressão linear multinível para analisar o efeito que cada uma das variáveis independentes criadas tem na competência medida pela variável R(POMP).

Será usada uma regressão linear para o modelo de nível 1, porque há apenas três pontos (três notas trimestrais) para cada indivíduo. A construção dos modelos seguirá as diretrizes apresentadas no segundo capítulo. Após chegar ao modelo que melhor explica os dados, apresentarei, usando recursos gráficos, os efeitos encontrados para cada preditor.

2.4.3. Análise multinível utilizando as repostas nas provas fechadas.

Em seguida, utilizarei um modelo de regressão linear multinível para analisar o efeito que cada uma das variáveis independentes criadas tem na competência medida pela variável R(ITENS).

Mais uma vez, a única opção é usar uma regressão linear como modelo de nível 1, pois há apenas três pontos (três provas) para cada indivíduo.

2.4.4. Comparação entre as duas análises

Após fazer as duas análises multinível e apresentar os resultados, farei uma comparação entre os resultados obtidos a partir das duas análises. Discutindo a diferença entre as competências medidas por cada uma delas, tentarei obter novas conclusões.

CAPÍTULO 3: RESULTADOS

Neste capítulo, apresentarei as análises da evolução das competências medidas pelas variáveis R(POMP) e R(ITENS), destacando e discutindo os resultados obtidos em cada uma. Em seguida, farei uma comparação entre os resultados das duas análises.

3.1. ANÁLISE EXPLORATÓRIA

3.1.1. Correlação entre R(POMP) e R(ITENS)

Antes de fazer a análise longitudinal por meio do modelo multinível, explorei os nossos dados para ganhar intuição sobre o comportamento de nossas variáveis dependentes, bem como de sua relação com algumas outras variáveis.

O primeiro ponto explorado foi a relação entre as nossas duas medidas de competência: as variáveis dependentes, R(POMP) e R(ITENS), considerando o conjunto das três ocasiões. O diagrama de dispersão abaixo nos dá uma idéia dessa relação.

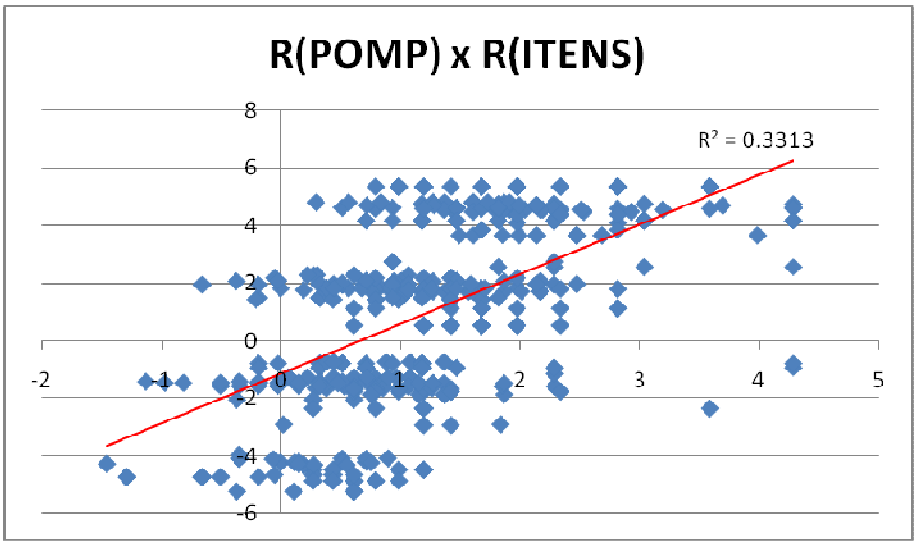

Gráfico 2: Dispersão - Variáveis R(POMP) e R(ITENS)

Pode-se observar que o coeficiente de correlação é $R = \sqrt{0{,}3313} = 0{,}5756$.

A estatística para testar a hipótese de que há correlação linear (conta a hipótese nula de que não há) é $t = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}}$, onde n é o número de pares da amostra.

Substituindo os valores (n = 3 x 145 = 435), temos t = 14,65, o que indica que há alguma correlação linear ($p<<0.01$).

Essa conclusão já era esperada, pois a variável R(ITENS) é obtida a partir da provas trimestrais e a R(POMP) é obtida a partir das notas trimestrais. Como o desempenho nas provas trimestrais contribui em grande parte para a nota do trimestre, é esperado que haja alguma relação entre as duas variáveis.

Apesar de o teste levar a rejeitar a hipótese nula (de que não há correlação linear), o coeficiente de correlação não é muito alto. Apenas essa breve exploração (apesar de útil para gerar intimidade com os dados) não me permite ir além dessas simples constatações.

3.1.2. Análise gráfica das trajetórias médias observadas

Antes de iniciar a análise multinível, fiz uma última exploração dos dados. Analisei, como antes, a evolução média das variáveis, mas, desta vez, separei os alunos por gênero e por turma.

Após explorar brevemente a relação entre as duas variáveis dependentes, resolvi explorar a evolução média dessas variáveis no tempo. Os gráficos abaixo mostram a média de cada uma dessas variáveis, entre todos os alunos, em função do tempo (em trimestres).

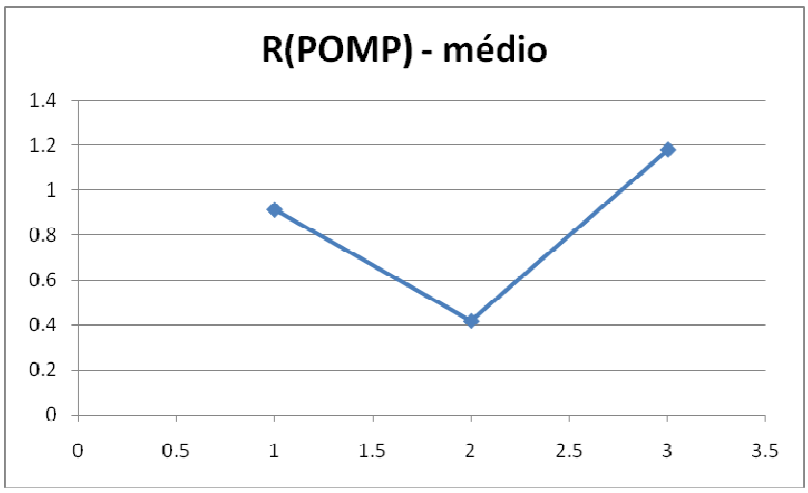

Gráfico 3: R(POMP) médio nas três ocasiões

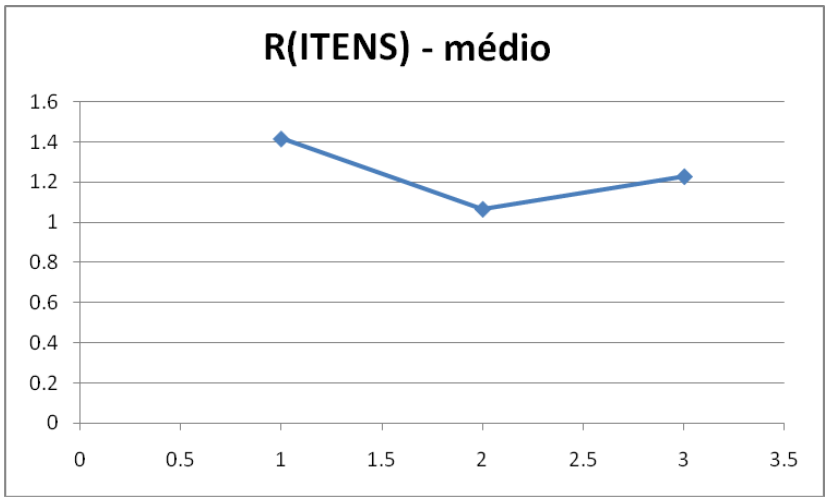

Gráfico 4: R(ITENS) médio nas três ocasiões

A partir da inspeção desse dois gráficos, pode-se observar que, em ambos, há um decréscimo médio do primeiro para o segundo trimestre e há um crescimento médio do segundo para o terceiro trimestre. Mas há também uma diferença: no gráfico do R(POMP) o ponto final é superior ao inicial, mas no gráfico do R(ITENS) o ponto final é inferior ao inicial).

Apenas as semelhanças e diferenças observadas não me permitem fazer nenhuma interpretação mais ousada, mas apenas reforçam a idéia de que as duas variáveis são medidas de construtos ligeiramente diferentes.

Evolução média por gênero

Os gráficos abaixo representam a evolução média das variáveis R(POMP) e R(ITENS), separadas por gênero.

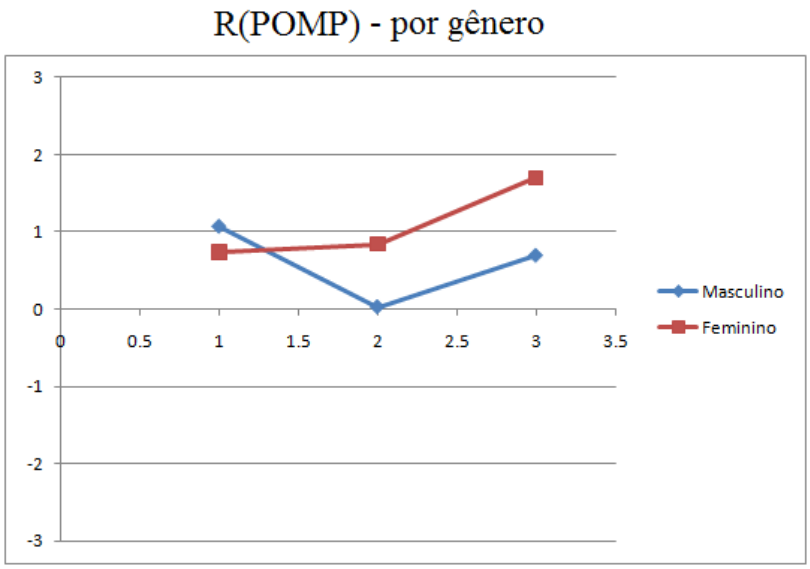

Gráfico 5: R(POMP) médio por gênero

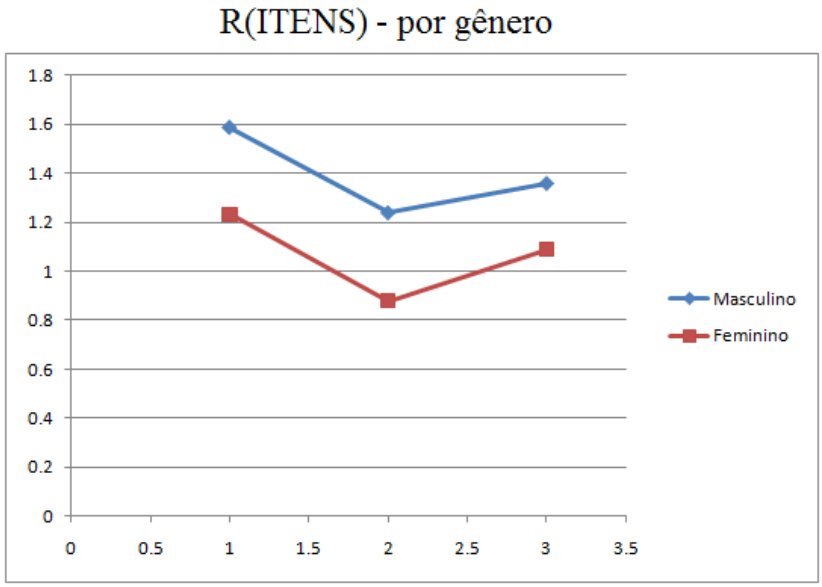

Gráfico 6: R(ITENS) médio por gênero

No gráfico do R(POMP), observa-se que no primeiro trimestre a média do R(POMP) é praticamente a mesma para meninos e meninas (a dos meninos ligeiramente mais alta). Observa-se também que, com o passar do tempo, a média das meninas aumenta, tronando-se maior que a dos meninos já no segundo trimestre e aumentando a diferença no terceiro.

Já no gráfico R(ITENS), ambos os grupos seguem trajetórias similares, mas com a competência dos meninos sendo sempre superior à das meninas.

Essa simples exploração sugere que algo substancialmente diferente deve haver entre as duas variáveis. Uma primeira reflexão me faz pensar que, como o R(POMP) é obtido a partir das notas trimestrais, ele é uma medida de uma competência mais ampla, envolvendo não só aspectos cognitivos, mas também esforço diário, assiduidade, cumprimento de regras e comportamento. Já o R(ITENS) foi obtido a partir do desempenho em provas com itens dicotômicos e presumo que a competência medida por essa variável envolva mais (mas não somente) aspectos cognitivos (conhecimentos e habilidades), que auxiliam na resolução de problemas de física.

Se essa interpretação está correta, significa que, na amostra, as meninas desenvolvem mais do que os meninos uma competência acadêmica ampla, mas a evolução de uma competência em física mais cognitiva ocorre de forma similar entre os grupos, mas com os meninos em um nível ligeiramente acima.

Evolução média por turma.

Os gráficos abaixo representam a evolução média das variáveis dependentes R(POMP) e R(ITENS) de cada turma.

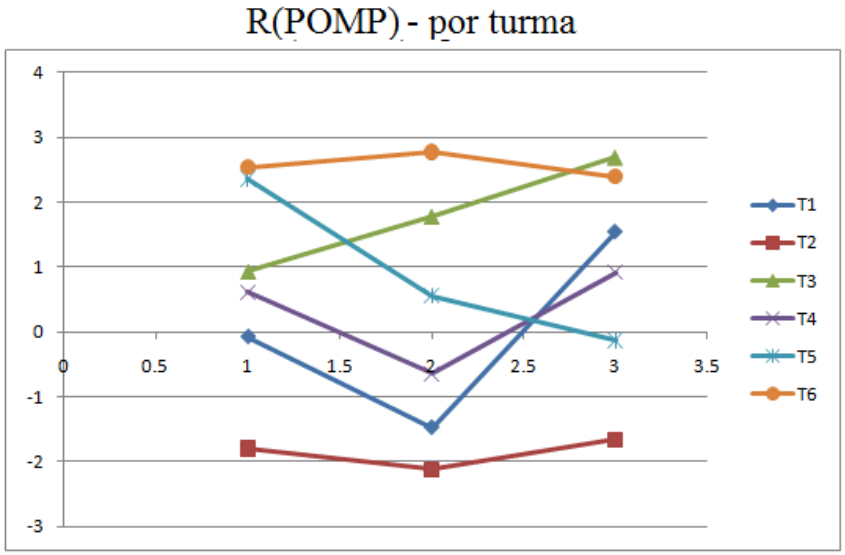

Gráfico 7: R(POMP) médio por turma

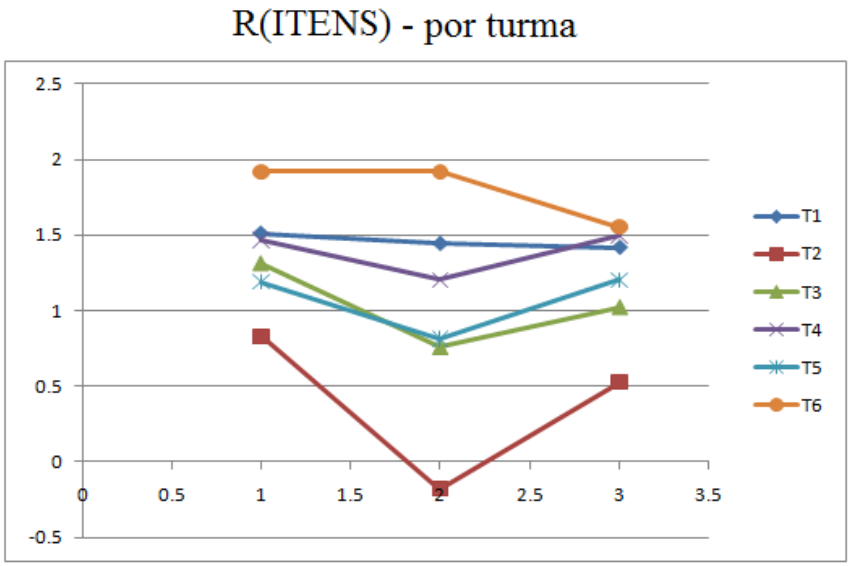

Gráfico 8: R(ITENS) médio por turma

Nos gráficos observa-se que: (i) para a variável R(POMP), as turmas T3 e T1 apresentam um grande crescimento; (ii) para a variável R(POMP), a turma T5 apresenta um grande decrescimento; (iii) para ambas as variáveis, a turma T2 tem uma trajetória que permanece abaixo da outras durante todo o processo.

É interessante notar que a turma T3, que apresente um crescimento acentuado e constante para o R(POMP), é uma turma formada por 22 meninas e apenas 5 meninos. Esse crescimento observado pode ter uma explicação comum ao crescimento observado para o R(POMP) do grupo feminino como um todo (esse crescimento da turma pode ser explicado pelo fato de a turma ser formada por meninas ou o crescimento das meninas pode ser explicado pelo crescimento dessa turma?).

3.2. ANÁLISE LONGITUDINAL DA VARIÁVEL R(POMP)

Essa análise longitudinal pretendeu investigar o efeito dos preditores criados na variável R(POMP). Utilizei os dados de 145 alunos (excluímos da análise dois sujeitos -

suj176 e suj240 – para usar os mesmos sujeitos usados na análise do R(ITENS)) e tratei os dados utilizando o software MLwiN.

O primeiro modelo que construí contém apenas um parâmetro para o intercepto que varia randomicamente entre os sujeitos (Modelo A):

$$R(POMP)_{ij} = B_0 + u_{0j} + e_{ij}$$

Nesse modelo, cada sujeito tem uma trajetória verdadeira plana, que difere da trajetória média, *R(POMP)* = B_0, por um fator u_{0j}. Os valores estimados para os parâmetros desse modelo encontram-se na tabela 1. Pode-se notar que a variância do coeficiente u_{0j} vale 4,691 e que a variância do coeficiente e_{ij} é 4,498. Isso indica que a variância da média do *R(POMP)* dos sujeitos em relação à média geral é muito próxima da variância do *R(POMP)* em diferentes ocasiões (do mesmo sujeito) em relação à média do sujeito. Em outras palavras: aproximadamente metade da variância ocorre entre os sujeitos (interindividual) e a outra metade ocorre "dentro" dos sujeitos (intraindividual).

Em seguida, acrescentei um termo para a variável temporal no modelo, obtendo o modelo B:

$$R(POMP)_{ij} = [B_0 + u_{0j}] + [B_1 + u_{1j}] \times TEMPO_{ij} + e_{ij}$$

Nesse modelo, cada pessoa tem uma trajetória verdadeira que difere da trajetória média. A diferença no intercepto (competência inicial) é dada pelo coeficiente u_{0j} e a diferença na inclinação é dada pelo coeficiente u_{1j}. Lembremos que o modelo assume que os

coeficientes u_{0j} e u_{1j} são ambos distribuídos normalmente com média zero, variâncias σ_0^2 e σ_1^2 (respectivamente) e covariância σ_{01}.

Acrescentar a variável TEMPO ao modelo fez com que a estatística desviância caísse de 2094,154 para 2084,742. Como o modelo B tem apenas um parâmetro a mais do que o modelo A, o teste qui-quadrado para a diferença das desviâncias indica que acrescentar a variável TEMPO proporcionou um ajuste significativamente melhor do modelo. Por isso, mantive essa variável no modelo.

Esse modelo, contendo apenas um parâmetro para o intercepto e outro para a inclinação, possibilita a visualização da evolução média da competência, medida pela variável R(POMP), no tempo. Essa evolução média está representada no gráfico abaixo.

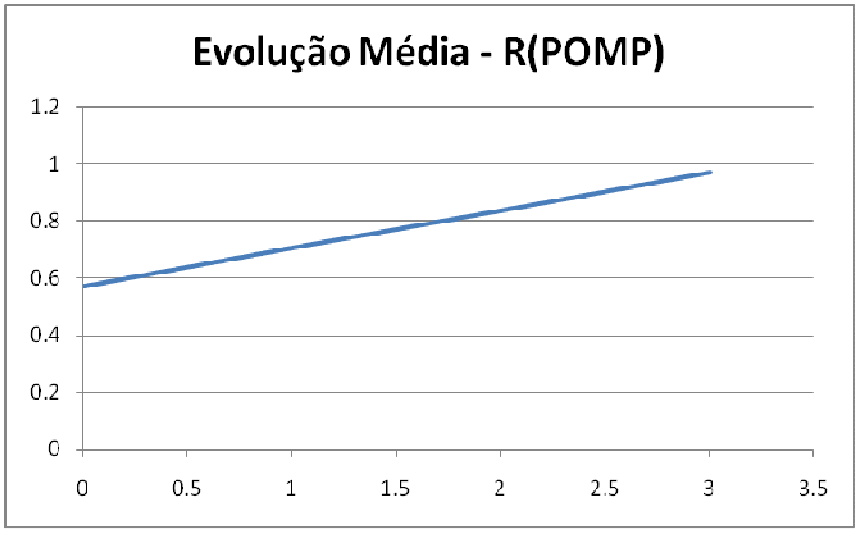

Gráfico 9: Evolução média prevista para a competência em física - R(POMP)

Após construir esse dois modelos, A e B, testei a inclusão, uma a uma, de todas as variáveis independentes criadas. Para cada nova variável introduzida, era avaliada a mudança na desviância e na variância não explicada (parâmetros randômicos). Quando a nova variável não contribuía para melhorar o ajuste do modelo, ela era descartada. Mas se ela contribuía para melhorar o ajuste, o novo modelo era armazenado. A tabela 1 traz um resumo de todos os

modelos que foram armazenados (ou seja, os que se ajustavam melhor que o anterior). O último deles, o modelo G, foi o que melhor se ajustou (melhor explicou a variância observada dos dados):

$$R(POMP)_{ij} = [B_0 + u_{0j}] + [B_1 + u_{1j}] \times TEMPO_{ij} + B_4 \times DPMmédio_j + B_5 \times DPMalto_j + B_6 \times E.Pai_j + B_7 \times T5_j + B_8 \times T3_j \times TEMPO_{ij} + B_9 \times T5_j \times TEMPO_{ij} + e_{ij}$$

R(POMP)	Modelo A		Modelo B		Modelo C		Modelo D		Modelo E		Modelo F		Modelo G	
	C	E.P.	C	E.P.	C	E.P.	C	E.P.	C	E.P.	Coefic.	E.P.	C	E.P.
Parte Fixa														
B_0	0.837	0.207	0.570	0.344	-1.467	0.365	-1.673	0.343	-2.241	0.392	-2.760	0.384	-2.682	0.377
B_1			0.134	0.138	0.134	0.138	0.134	0.138	0.328	0.149	0.357	0.154	0.347	0.154
B_2					2.534	0.313	0.296	0.429	0.617	0.485	0.424	0.440		
B_3					5.235	0.428	1.822	0.705	1.813	0.762	1.404	0.692		
B_4							2.775	0.427	2.571	0.490	2.641	0.440	3.002	0.292
B_5							3.744	0.673	3.502	0.730	4.119	0.666	5.265	0.348
B_6									0.685	0.298	0.647	0.268	0.620	0.267
B_7											0.686	0.141	0.723	0.142
B_8											3.979	0.915	3.951	0.918
B_9											-1.360	0.411	-1.351	0.411
Parte randómica														
σ_{u0}^2	4.691	0.738	9.123	2.219	6.242	1.918	5.112	1.803	4.648	1.996	2.817	1.794	2.894	1.803
σ_{u01}			-2.072	0.835	-2.165	0.806	-2.063	0.787	-1.686	0.850	-1.022	0.772	-1.028	0.773
σ_{u1}^2			1.053	0.382	1.054	0.382	1.053	0.382	0.778	0.404	0.467	0.370	0.470	0.371
σ_e^2	4.498	0.374	3.427	0.402	3.427	0.402	3.427	0.402	3.423	0.457	3.423	0.457	3.423	0.457
Desviância	2094.154		2084.742		1976.608		1935.783		1479.905		1440.910		1445.074	

Tabela 1: Modelos construídos para a variável R(POMP)

Os gráficos abaixo ilustram o efeito na evolução da competência, previsto pelo modelo para cada uma das variáveis, depois de controlados os efeitos para as outras variáveis.

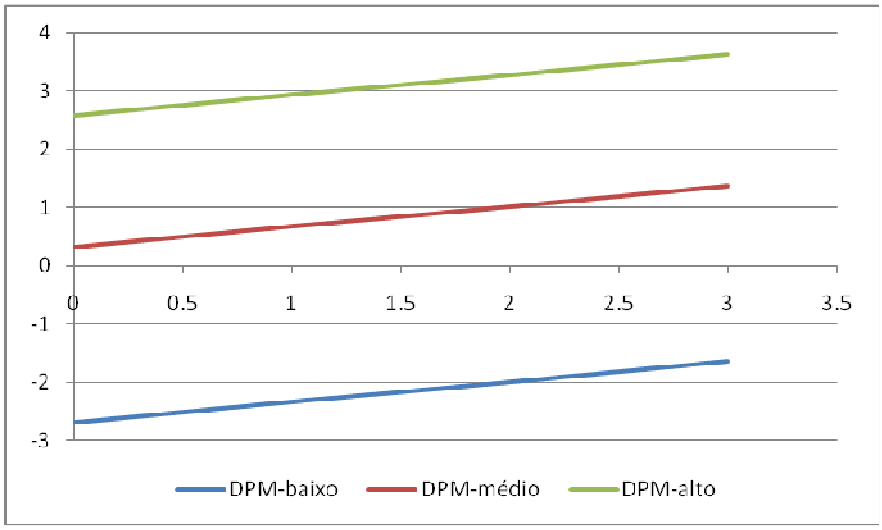

Gráfico 10: Evolução média da competência - variável R(POMP) - por grupos de desempenho prévio em matemática - com todas as outras variáveis assumindo valor zero

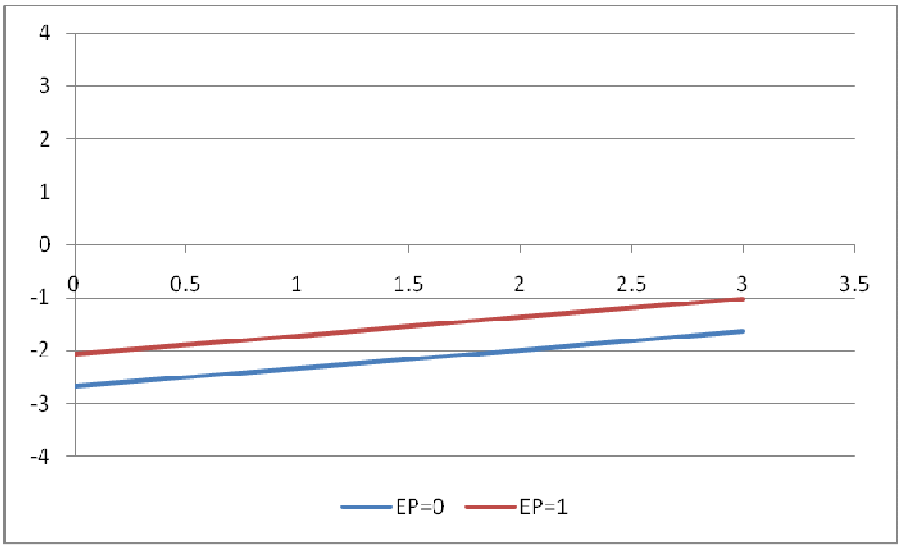

Gráfico 11: Evolução média da competência - variável R(POMP) - por grupos de escolarização do pai - com todas as outras variáveis assumindo valor zero

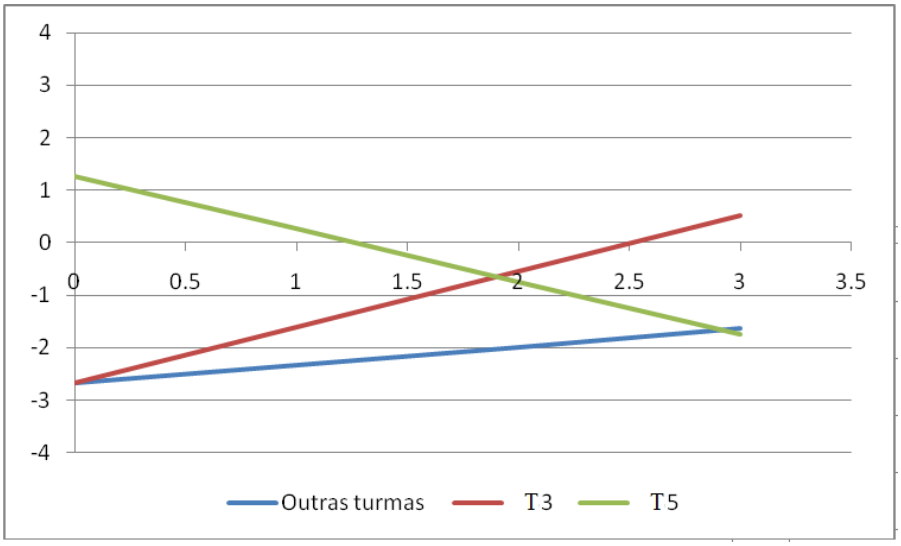

Gráfico 12: Evolução média da competência - variável R(POMP) por turma - todas as outras variáveis assumindo valor zero

A partir das estimativas para os coeficientes do modelo ajustado (na tabela 1) e dos gráficos apresentados, pode-se observar que:

i- As trajetórias verdadeiras individuais diferem entre si, tanto no intercepto ($\sigma_{u0}^2>0$), quanto na inclinação ($\sigma_{u1}^2>0$).

ii- A trajetória média da competência apresenta um crescimento durante o ano.

iii- O grupo de desempenho prévio em matemática é o fator que mais influencia na trajetória da competência na terceira série, sendo a trajetória média mais alta para o grupo de alunos com DPM alto, e a mais baixa para o grupo de alunos com DPM baixo (mas a inclinação é a mesma).

iv- A trajetória média da competência não difere entre os grupos de desempenho prévio de física (nem no intercepto, nem na inclinação).

v- A trajetória média da competência prevista para o grupo com EP=1 (pai concluiu o Ensino Superior) é mais alta, mas tem a mesma inclinação que a trajetória média prevista para o grupo com EP=0 (pai não concluiu o Ensino Superior).

vi- Após controlar o efeito de todas as variáveis independentes criadas, apenas algumas turmas ainda apresentam diferenças nas suas trajetórias médias previstas para a competência. A turma T3 tem um crescimento maior que as outras. A turma T5 começa com um alto valor de competência média, mas tem um grande decaimento.

vii- A renda familiar, a escolarização da mãe e o professor não apresentaram efeitos significativos, controlados os efeitos das outras variáveis.

O crescimento da competência durante o ano é algo esperado, já que se trata de um processo de aprendizagem. No entanto, deve-se lembrar que a competência medida pela variável R(POMP) é um composto de componentes cognitivos e fatores relacionados a esforço, engajamento, entre outras coisas, já que a variável é obtida a partir da nota trimestral. Portanto, não se pode ser muito assertivo ao afirmar que as habilidades e conhecimentos no domínio da física estão aumentando.

A observação de que o desempenho prévio em matemática tem um grande efeito sobre a trajetória da competência pode se dever às características do curso de física na terceira série. O material didático impresso utilizado, de autoria do professor efetivo e coordenador da série, e as atividades propostas, exigiam conhecimento de matemática, mas não faziam qualquer revisão nem ensinavam os conhecimentos de matemática necessários, que eram usados como se já fossem bem conhecidos.

A não detecção de um efeito significativo para o desempenho prévio em física (após controlar o efeito do desempenho prévio em matemática) parece indicar que, além das habilidades e conhecimentos que a física tem em comum com a matemática, não há uma contribuição adicional da competência em física nas séries anteriores na trajetória da terceira série. Isso pode se dever à diferença de tratamento da física na terceira série em relação às séries anteriores.

Já o efeito encontrado para a escolarização do pai (ou da mãe) poderia ser, de certa forma, esperado. Esse efeito é, de fato, coerente com resultados de outras pesquisas (CATSAMBIS, 1998; JOHNSON, 2009). Isso pode refletir uma suposta expectativa que os pais criam (e que o aluno acaba absorvendo) sobre a trajetória acadêmica dos estudantes. Embora não seja óbvia, parece ser razoável a hipótese de que, de uma forma geral, pais com Ensino Superior completo valorizem, em média, mais a vida acadêmica e por isso se tornem mais exigentes em relação ao desempenho de seus filhos na escola.

Não sou capaz de explicar as diferenças nas trajetórias encontradas para as turmas T3 e T5. Deve-se notar, no entanto, que essas diferenças aparecem mesmo após controlarmos os efeitos de todas as outras variáveis e que, dessa forma, possivelmente são devidas a características próprias das turmas. Nesse caso, posso apenas destacar algumas características específicas dessas turmas. No caso da turma T3, lembremos que é única turma com uma quantidade de meninas (22) muito maior que de meninos (5). Isso, porém, não deve ser uma explicação, pois, se fosse, haveria um efeito para a variável GÊNERO. Além disso, essa é a única turma cujos alunos fazem o curso técnico de “Patologia Clínica”. Quanto à turma T5, não consegui destacar nenhuma particularidade em relação às outras.

3.3. ANÁLISE MULTINÍVEL DA VARIÁVEL R(ITENS)

Essa análise longitudinal pretendeu investigar evolução da competência, medida pela variável R(ITENS), bem como o efeito dos preditores criados nessa evolução. Foram utilizados os dados de 145 alunos (excluí da análise dois sujeitos - suj176 e suj240 - que não haviam feito todas as três provas) e tratei os dados utilizando o software MLwiN.

O primeiro modelo que construí para a variável R(ITENS) é um modelo contendo apenas um parâmetro para o intercepto que varia randomicamente entre os sujeitos (Modelo A):

$$R(ITENS)_{ij} = B_0 + u_{0j} + e_{ij}$$

Os valores estimados para os parâmetros desse modelo encontram-se na tabela 2. Dela, pode-se perceber que o valor médio da competência, medida pela variável R(ITENS), para todos os sujeitos e ocasiões é 1,237(0,069). A variância entre as médias dos sujeitos é 0,522(0,082) e a variância dentro dos sujeitos (variância das medidas em relação às médias dos sujeitos) é 0,507(0,042). Isso significa que a variância intraindividual e a variância interindividual correspondem a aproximadamente metade da variância total.

Em seguida, acrescentei a variável temporal no modelo, obtendo o modelo B:

$$R(ITENS)_{ij} = B_0 + B_1 \times TEMPO_{ij} + u_{0j} + e_{ij}$$

Diferentemente do modelo B para a variável R(POMP), esse modelo não tem um coeficiente randômico para a variável TEMPO (um modelo que inclui esse o coeficiente randômico não se ajustou bem). Isso significa que o modelo prevê trajetórias verdadeiras individuais, todas com a mesma inclinação, B_1 (embora com interceptos diferentes).

Esse modelo, contendo apenas um parâmetro para o intercepto e outro para a inclinação, possibilita a visualização da evolução média da competência, medida variável R(POMP), no tempo. Essa evolução média está representada no gráfico abaixo.

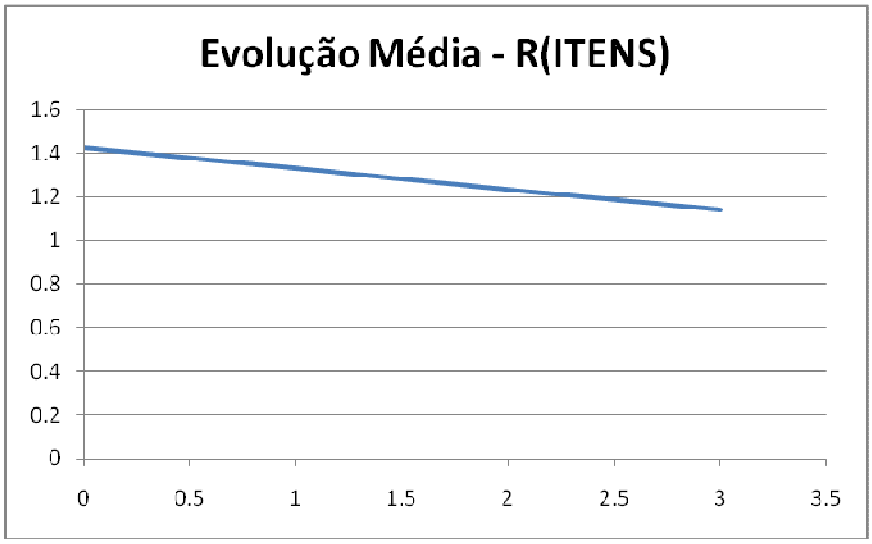

Gráfico 13: Evolução média da competência em física - variável R(ITENS)

Seguindo um processo semelhante ao descrito para a variável R(POMP), fui construindo modelos, acrescentando novas variáveis. Um resumo dessa história (contendo apenas os modelos que melhoraram o ajuste) pode ser encontrado na tabela 2. O modelo que melhor se ajustou aos dados foi o modelo G:

$$R(ITENS)_{ij} = [B_0 + u_{0j}] + B_1 \times TEMPO_{ij} + B_2 \times DPFmédio_j +$$
$$B_3 \times DPFalto_j + B_4 \times DPMmédio_j + B_5 \times DPMalto_j + B_6 \times GÊNERO_j +$$
$$B_7 \times E.Mae_j + B_8 \times PROF1_j \times TEMPO_{ij} + e_{ij}$$

R(ITENS)	Modelo A		Modelo B		Modelo C		Modelo D		Modelo E		Modelo F		Modelo G	
	C	E.P.	C	E.P.	C	E.P.	C	E.P.	C	E.P.	C	E.P.	C	E.P.
Parte Fixa														
B_0	1.237	0.069	1.425	0.108	0.953	0.122	0.882	0.116	1.060	0.119	0.895	0.130	0.852	0.130
B_1			-0.094	0.041	-0.094	0.041	-0.094	0.041	-0.094	0.041	-0.080	0.044	-0.110	0.047
B_2					0.474	0.119	-0.284	0.168	-0.144	0.160	-0.229	0.165	-0.214	0.163
B_3					1.554	0.163	0.346	0.276	0.477	0.259	0.304	0.263	0.318	0.259
B_4							0.935	0.167	0.808	0.158	0.788	0.165	0.767	0.163
B_5							1.330	0.263	1.291	0.246	1.415	0.248	1.428	0.244
B_6									-0.424	0.093	-0.328	0.098	-0.247	0.104
B_7											0.254	0.098	0.247	0.097
B_8													0.102	0.050
Parte Randómica														
σ_{u0}^2	0.522	0.082	0.525	0.082	0.253	0.052	0.172	0.042	0.130	0.037	0.108	0.037	0.099	0.036
σ_e^2	0.507	0.042	0.499	0.041	0.499	0.041	0.499	0.041	0.499	0.041	0.437	0.041	0.437	0.041
Desvi â ncia	1143.502		1138.442		1067.584		1035.003		1015.630		737.992		733.940	

Tabela 2: Modelos construídos para a variável R(ITENS)

Os gráficos abaixo ilustram o efeito na evolução da competência em física, previsto pelo modelo para cada uma das variáveis, depois de controlados os efeitos para as outras variáveis.

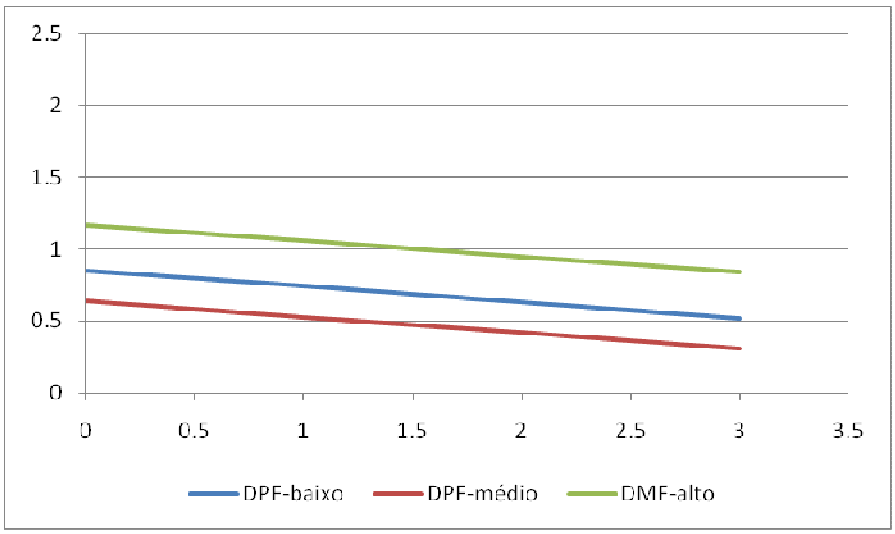

Gráfico 14: Evolução média da competência - variável R(ITENS) - por grupo de desempenho prévio em física - todas as outras variáveis assumindo valor zero

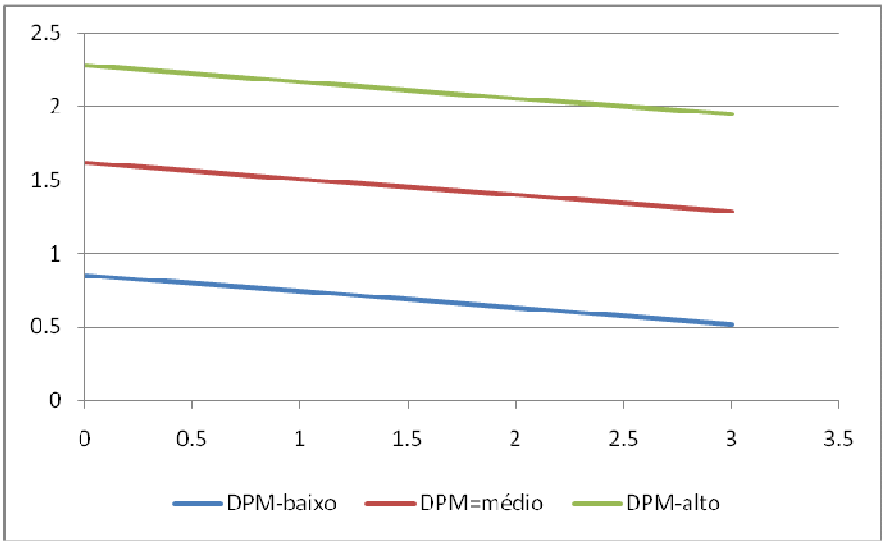

Gráfico 15: Evolução média da competência - variável R(ITENS) - por grupo de desempenho prévio em matemática - todas as outras variáveis assumindo valor zero

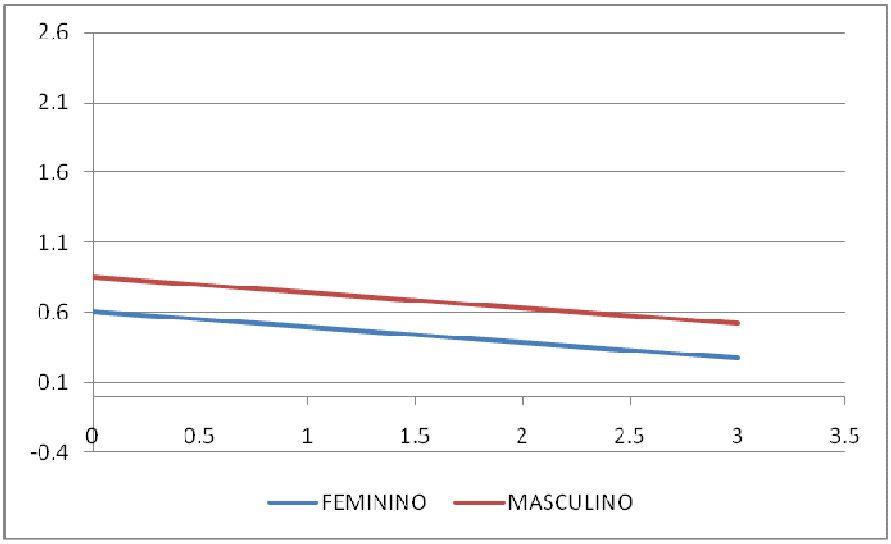

Gráfico 16: Evolução média da competência - variável R(ITENS) - por gênero - todas as outras variáveis assumindo valor zero

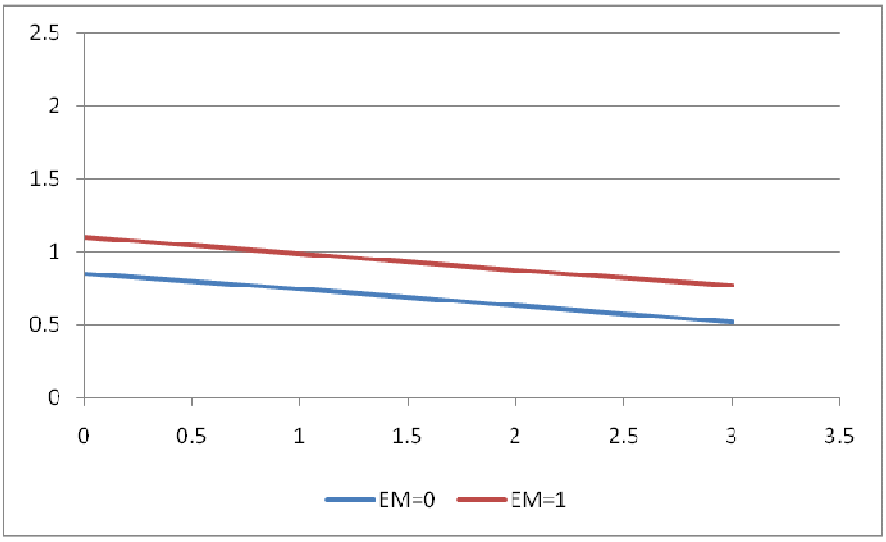

Gráfico 17: Evolução média da competência - variável R(ITENS) - por grupos de escolarização da mãe - todas as outras variáveis assumindo valor zero

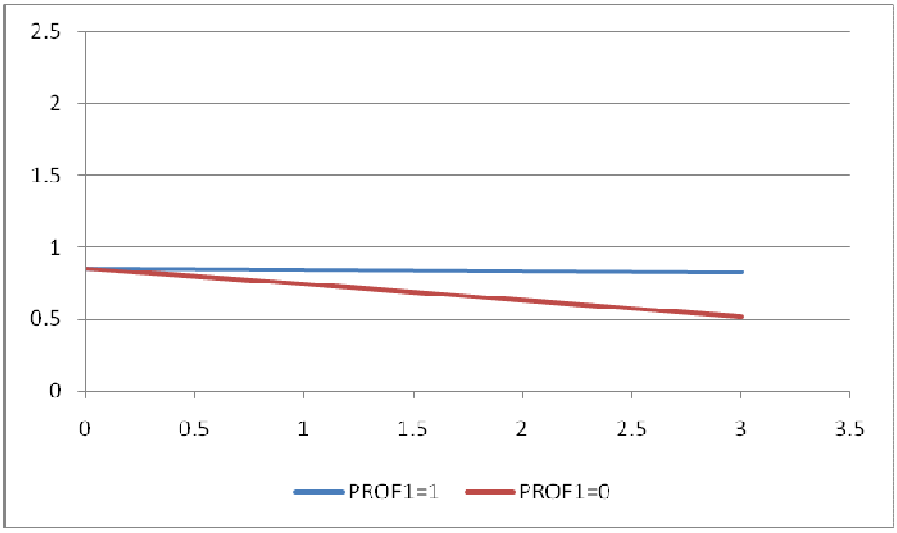

Gráfico 18: Evolução média da competência - variável R(ITENS) - por professor - todas as outras variáveis assumindo valor zero

A partir das estimativas apresentadas para esse modelo na tabela 2 e dos gráficos apresentados, pode-se observar que:

i- Os interceptos das trajetórias verdadeiras individuais têm uma variância significativamente diferente de zero, ou seja, os alunos não começam a terceira série com o mesmo nível de competência.

ii- Não há variância para o parâmetro do intercepto, ou seja, o modelo prevê a mesma inclinação para todos os alunos que pertençam a um mesmo grupo.

iii- A trajetória verdadeira média prevista é ligeiramente (mas significativamente) decrescente.

iv- Os diferentes grupos de desempenho prévio em física diferem em seus interceptos médios (mas não na inclinação), mas a única dessas diferenças que é estatisticamente significativa é a do intercepto do grupo “DPF-alto” em ralação ao grupo “DPF-médio”.

v- O desempenho prévio de matemática é a variável que exerce o maior efeito no intercepto. O grupo DPM-alto é o que tem o maior intercepto médio previsto, seguido pelo grupo DPM-médio.

vi- O intercepto médio (competência no tempo zero) previsto é maior para os meninos que para as meninas, mas a inclinação (variação da competência com o tempo) não é diferente.

vii- A média do intercepto para os alunos cuja mãe completou o Ensino Superior é maior que a média para os demais alunos (mas o mesmo não ocorre para a inclinação).

viii- As variáveis relativas à renda familiar, professor e turma não têm efeito nem no intercepto, nem na inclinação.

A observação de que o desempenho prévio em matemática tem um efeito muito maior que o desempenho prévio de física no intercepto (que é o valor da competência no início da terceira série) pode ter uma explicação semelhante ao caso da variável R(POMP). Ou seja, pode ser devida à forte exigência de conhecimentos e habilidades de matemática nas provas, sem que seja dado suporte para esses conhecimentos durante a terceira série.

O efeito do desempenho prévio em física deve ser visto com cuidado, já que nenhum dos coeficientes é estatisticamente significativo.

A observação de que os meninos têm uma competência média inicial maior do que as meninas e que essa diferença se mantém durante o curso é condizente com resultados relatados na literatura para a diferença de desempenho em física (MULLER *et al*, 2001; BYRNES e MILLER, 2007; LAWRENZ *et al*, 2009, GRIGG *et al*, 2006), mas não sou capaz de explicá-los.

Embora tenha sido encontrado um efeito do professor 1 na inclinação, uma reflexão mais aprofundada me aconselha certa cautela ao interpretar esse efeito. Isso porque a variável professor não tem a estabilidade que pode parecer, à primeira vista. Um professor pode ter um papel diferente para cada turma. Dizer que o "professor 1" tem um efeito positivo na trajetória é perigoso porque não se pode afirmar que esse efeito ocorreria se ele lecionasse para outra turma. Talvez só faça sentido falar do efeito de uma combinação professor-turma, mas não sobre um efeito de um professor, como se esse efeito fosse transferível.

Um último resultado encontrado é ausência de um "efeito turma". Esse é, de certa forma, um resultado inesperado, pois as turmas são definidas a partir do curso técnico escolhido pelos alunos. Dessa forma, elas devem refletir certo vocacionamento e pesquisas anteriores (COELHO e BORGES, 2010) indicam que o vocacionamento influencia no aprendizado de física.

Sem dúvida, a observação mais difícil de explicar é a de que a competência medida pela variável R(ITENS) cai (pouco, mas significativamente) durante o curso. O que parece é que os alunos saem da terceira série com uma competência ligeiramente menor que entraram. Pode ser tentador pensar que isso ocorreu por causa de um decréscimo de engajamento durante o curso, mas essa hipótese não resiste à observação de que a variável R(POMP), que tem uma componente muito maior de engajamento, aumenta durante o ano. Essa observação,

tão relevante e difícil de entender, pode ter sua explicação em certa inadequação do desenho metodológico da pesquisa. Dedicarei uma seção do capítulo “Discussões e conclusões” apenas à discussão dessa questão.

3.4. COMPARAÇÃO ENTRE AS DUAS ANÁLISES.

A última parte da análise consiste na comparação das duas análises multinível (com o R(POMP) e com o R(ITENS). Pode-se verificar que há algumas semelhanças, mas há também diferenças fundamentais.

A primeira semelhança é a grande influência do desempenho prévio em matemática no intercepto das trajetórias (mas não na inclinação). O grupo de alunos com desempenho prévio em matemática “alto” tem uma trajetória média superior ao grupo com desempenho prévio em matemática “médio” e esse, por sua vez, tem uma trajetória média superior ao grupo com desempenho prévio em matemática “baixo”.

Outra semelhança é o fato do desempenho prévio em física não influenciar (caso do R(ITENS) ou quase não influenciar (caso do R(POMP)) a trajetória da competência em física na terceira série.

A existência dessas duas semelhanças corrobora com a explicação para a detecção de um efeito do desempenho prévio em matemática ser muito maior que o de física: a física na terceira série exige mais conhecimentos de matemática que nas séries anteriores, e esse conhecimento é tratado como supostamente sabido.

Há ainda outra semelhança: a renda familiar não influencia nenhuma das trajetórias.

Vamos agora às diferenças entre os resultados das duas análises.

A primeira (e talvez mais fundamental) diferença é que a trajetória média prevista para a variável R(POMP), apresenta um crescimento durante o ano, enquanto a variável R(ITENS) apresenta um decaimento. Podemos tentar entender essa diferença pensando em quais pontos as competências medidas pelas variáveis diferem entre si. Ora, a variável R(ITENS) foi obtida a partir de respostas a itens dicotômicos, enquanto a variável R(POMP) foi obtida a partir de notas trimestrais. Essas notas trimestrais eram distribuídas em diversas atividades (já mencionadas). Por isso, a competência medida pela variável R(POMP) envolve, além de habilidades e conhecimentos cognitivos (que são as principais componentes do R(ITENS)), fortes componentes de esforço, engajamento, freqüência e participação nas aulas. Então é plausível pensar que, se a variável R(POMP) aumenta e R(ITENS) diminui, esses outros aspectos mencionados (esforço, engajamento, *etc.*) é o que está contribuindo para o aumento da variável R(POMP). Ou seja, os alunos estão, em média, aumentando seu engajamento, esforço e participação nas aulas.

Outra diferença importante é o efeito da variável GÊNERO. Os meninos apresentam maior competência que as meninas quando a análise é feita como o R(ITENS), mas não há diferença de competência se a análise é feita com o R(POMP). Mais uma vez, tentarei explicar essa diferença a partir das diferenças das próprias variáveis dependentes. Como já mencionei, a variável R(ITENS) está mais relacionada com aspectos cognitivos, enquanto o R(POMP) envolve componentes de esforço e engajamento. Então é plausível pensar que as meninas, em média, estiveram mais engajadas e se esforçaram mais que os meninos. Isso acabou compensando a menor competência cognitiva (detectada no R(ITENS)) e fez com que não houvesse diferença de gênero no R(POMP).

Uma terceira diferença é que, para a variável R(ITENS), não foi detectada diferença entre as trajetórias médias das turmas, enquanto que para o R(POMP) houve diferença. Mais

uma vez, essa diferença deve estar ocorrendo nas componentes de engajamento e esforço que compõem o R(POMP).

Outra diferença aparente diz respeito à escolarização dos pais. Na análise feita com a variável R(POMP) foi encontrado um efeito para a escolarização do pai, enquanto que na variável R(ITENS) o efeito encontrado foi para a escolarização da mãe. Isto sugere certo cuidado ao interpretar tais diferenças. Na verdade, quando o efeito da escolarização do pai não é controlado, a escolarização da mãe apresenta efeito sobre o R(POMP). Da mesma forma, quando o efeito da escolarização da mãe não é controlado, a escolarização do pai tem efeito sobre o R(ITENS). A questão é que, na análise do R(POMP), o efeito da escolarização do pai é maior que o da escolarização da mãe e, somente após controlar o efeito da escolarização do pai, a da mãe deixa de influenciar. Isso pode ser explicado pelo fato de mais de 70% dos alunos terem os dois pais no mesmo grupo de escolarização (E.Pai=1 e E.Mae=1 ou E.Pai=0 e E.Mae=0). Isso indica que as duas variáveis estão correlacionadas e seus efeitos misturados. O que pode ter determinado qual das duas teve maior influência em cada variável dependente pode não ter muita importância. Então, o que seria, em princípio, uma diferença, pode ser visto como uma semelhança: em ambas as variáveis, alunos cujos pais (pai, mãe, ou ambos) concluíram Ensino Superior têm trajetória média superior à trajetória média dos outros alunos. Esse também é um resultado coerente com a literatura (CATSAMBIS, 1998; JOHNSON, 2009).

Não considero relevante o aparecimento da variável “Prof1” na análise do R(ITENS), pelos motivos já discutidos.

CAPÍTULO 4: DISCUSSÕES E CONCLUSÕES

4.1. DECAIMENTO DA VARIÁVEL R(ITENS)

Sem dúvida, o ponto mais desafiador desta pesquisa é entender e explicar o decaimento da competência detectado na análise da variável R(ITENS). Para tentar entender esse decaimento, vou analisar algumas características do desenho metodológico da pesquisa.

Primeiramente, lembremos que a variável R(ITENS) foi obtida a partir das respostas aos itens das provas trimestrais. Como o conteúdo cobrado nas trimestrais era acumulativo, a amplitude do domínio que estava sendo avaliado foi aumentando. Isso pode ter criado um efeito de multidimensionalidade na medida. Ou seja, pode ser que o que foi medido em cada ocasião não seja exatamente a mesma coisa.

Para investigar se isso pode estar de fato ocorrendo, criei outra variável dependente, de forma semelhante à que criei a variável R(ITENS), mas usando apenas questões do conteúdo de mecânica (comum às três provas). Chamei essa variável de R(MECANICA). Por usar apenas questões do mesmo assunto nas três ocasiões, o R(MECANICA) não deve apresentar o efeito do aumento de abrangência e complexidade que pode ser o causador do decaimento da variável R(ITENS). De fato, em média, essa variável não apresentou decaimento, mas crescimento, durante o curso, como indica o gráfico abaixo:

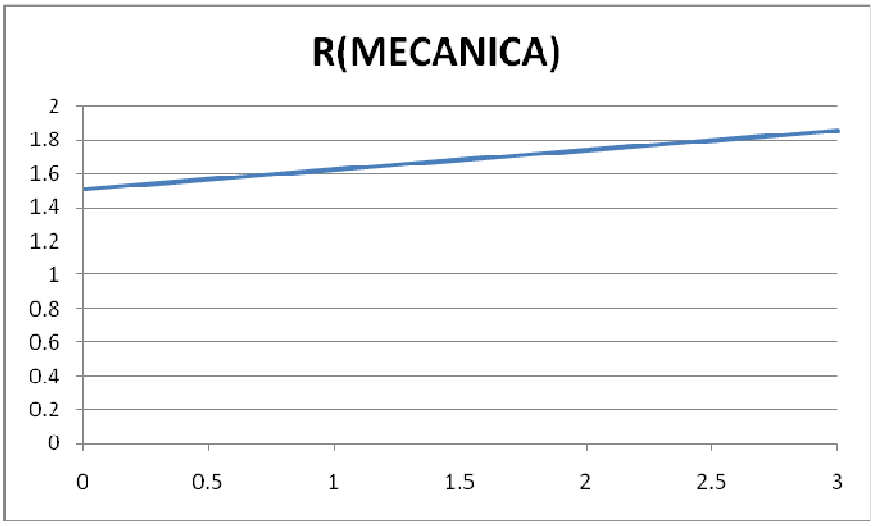

Gráfico 19: Evolução média da variável R(MECANICA)

Nesse caso, a suposição de que se está medindo a mesma coisa nas três ocasiões é menos problemática. No entanto, esse R(MECANICA) representa a competência em mecânica, não a competência em física, como um todo.

Para obter uma competência em física, como um todo, deveriam ser usados itens de todos os campos da física, mas sem que houvesse um aumento da abrangência com o tempo. Mas não tinha dados para isso, já que cada prova cobria apenas o conteúdo estudado até o momento de sua aplicação.

Também não posso supor que a competência em física evolua da mesma forma que a competência em mecânica, pois cada subdomínio pode evoluir de uma forma diferente, de acordo com o que está sendo trabalhado em sala de aula. No presente caso, penso que a competência em física deveria aumentar mais que a competência em mecânica, pelo raciocínio exposto abaixo:

Suponhamos a seguinte situação hipotética:

Um grupo de alunos é submetido a dois testes: um teste de mecânica e um teste de circuitos elétricos. Após serem submetidos a esses testes, os alunos participam de um curso de circuitos elétricos. No final do curso, os alunos são novamente submetidos a dois testes,

um de mecânica e outro de circuitos elétricos (ambos possíveis de serem equalizados em relação aos primeiros testes). O esperado pelo bom senso e indicado por pesquisas (RECKASE, 2004; SAYRE e HCKLER, 2009) é que haverá um maior aumento de competência no domínio em que houve instrução (no caso, circuitos). Pode ser que a competência em mecânica aumente, permaneça constante ou até diminua. Mas espera-se que o aumento na competência em circuito seja maior que em mecânica.

No caso desta pesquisa, ocorreu algo muito parecido. A diferença é que não havia um teste sobre o domínio estudado no momento anterior ao seu estudo. Por isso, não haveria forma de captar um possível aumento da competência onde ele deveria ser maior.

Apesar dessas ponderações, o teste de dimensionalidade feito com a terceira prova, Trimestral3, que incluía todos os conteúdos do ano, não acusou um efeito de multidimensionalidade. Além disso, a verificação do funcionamento dos itens para o conjunto das três provas também sugeriu um possível tratamento unidimensional. Por esses motivos, procedi com a análise da forma descrita na dissertação.

Não tenho uma explicação definitiva para isso, mas me parece que para examinar essa questão não se deve tratar a competência apenas como uma soma de “habilidades” e “conhecimentos”, mas tratar separadamente vários tipos de habilidades e vários tipos de conhecimento. Essa é, claramente, uma abordagem multidimensional e talvez um modelo multidimensional, com os dados apropriados, fosse necessário para uma investigação adequada.

Por limitações impostas pelos dados, não foi possível fazer um tratamento multidimensional. Mas especularei um pouco, a fim de jogar alguma luz sobre a questão.

Reckase (2009) afirma que o número de dimensões avaliadas por um teste depende não só do próprio teste, mas também da população para quem se aplica esse teste. Se a

população não tiver variabilidade suficiente em uma das dimensões que o teste mede, ele não será capaz de avaliar essa dimensão.

Pode ser que, no presente caso, as populações nas três ocasiões de medida não sejam as mesmas (apesar de serem os mesmos alunos) e que o grau de variabilidade em alguma das dimensões dos testes tenha mudado de uma ocasião para a outra.

Consideremos, por exemplo, uma dimensão bastante estreita como o "domínio do vocabulário usado em eletricidade" (esse pode ser um componente do conhecimento declarativo). Pode ser que na ocasião da terceira trimestral essa dimensão não tenha variabilidade suficiente para ser detectada no teste de dimensionalidade (todos os alunos teriam o vocabulário de eletricidade razoavelmente desenvolvido). No entanto, pode ser que se aplicássemos essa prova à população (imaginária) contendo todos os sujeitos em todas as ocasiões, essa população teria uma grande variabilidade nessa dimensão (já que alguns já estudaram os conteúdos de circuitos elétricos e outros não). Dessa forma, a prova poderia ser unidimensional para cada ocasião separadamente, mas multidimensional para um estudo longitudinal, usando as três ocasiões.

Essas são apenas especulações vagas e sem sólida fundamentação empírica, mas podem ajudar a intuir algumas possíveis soluções em uma possível pesquisa futura que lide com a questão.

É importante ressaltar que, se esse efeito ocorreu para o R(ITENS), deve ocorrer também para o R(POMP). No entanto, entendo que isso não invalida a comparação entre as duas análises. Além disso, o fato de ter encontrado resultados consistentes (muitos em acordo com a literatura) sugere que o viés encontrado para a inclinação das trajetórias também não invalida os efeitos encontrados para as diferenças nos interceptos.

Mas resta ainda uma questão: é possível encontrar algum sentido para a queda que identificamos? A limitação metodológica discutida acima talvez tenha comprometido

completamente a interpretação desse decaimento. Porém, não temos condições de especular mais sobre o tema pela falta de informações adicionais.

4.2. O USO DE AVALIAÇÔES ESCOLARES EM PESQUISAS EDUCACIONAIS

Após a análise dos resultados, volto à pergunta inicial: é possível usar avaliações escolares ordinárias para estudar a evolução da competência em física?

O fato de ter obtido vários resultados consistentes (já discutidos) nos sugere que a reposta é sim. No entanto, há várias ressalvas.

4.2.1. Ressalvas

I – Vagueza da conceituação de competência

O uso de notas trimestrais leva a uma definição de competência que não envolve apenas aspectos cognitivos, mas também fatores de engajamento, esforço, comportamento, entre outros. Para um estudo mais detalhado das componentes dessa “competência”, podem-se usar notas obtidas em diferentes tipos de avaliação: provas fechadas, provas abertas, exercícios, conceito e participação, entre outros. Algumas das conclusões (como um maior engajamento das meninas, por exemplo) só puderam ser obtidas pela comparação entre análises usando dois tipos de avaliações diferentes. Entendo que quanto mais tipos diferentes de avaliações, mais potencial terá a pesquisa.

II - Tratamento multidimensional

O problema do decaimento da variável R(ITENS) parece aconselhar a fazer um tratamento multidimensional (pelo menos para estudos longitudinais de mudança). O uso de respostas a questões abertas (embora esses sejam dados difíceis de obter) poderia tornar possível essa análise das muitas dimensões da competência.

Para contornar a multidimensionalidade, que parece surgir do aumento da abrangência do conteúdo, devemos: ou nos restringir a análise da evolução da competência em um conteúdo restrito como fiz com o R(MECANICA); ou tratarmos as dimensões cognitivas (vários tipos de habilidades e conhecimentos) que não dependessem fortemente do conteúdo (mas não sei se isso poderia ser chamado de competência).

Outra opção seria ter um desenho de coleta de dados que permitisse medir a competência em certos conteúdos antes e depois que esses fossem estudados, mas essa estrutura de avaliação contraria o bom senso da prática educativa (a não ser em casos de currículo em espiral). Ao se montar uma seqüência de testes com essa estrutura de coleta, sairíamos do domínio das avaliações ordinárias escolares e entraríamos no domínio dos testes normalmente usados em pesquisas, com todos os seus problemas.

III- Dados equalizáveis, métrica estável

Como foi discutido no capítulo “REFERENCIAIS TEÓRICOS”, um estudo longitudinal exige uma métrica que seja estável no tempo. Nesse caso, é importante ressaltar que não se deve usar a simples nota (ou conceito) obtida em provas ou no trimestre para a análise longitudinal, mas deve-se buscar um jeito de criar uma medida em uma escala que apresente forte estabilidade temporal (no presente caso, usei o modelo Rasch).

Além disso, as medidas obtidas em diferentes ocasiões devem poder ser comparadas umas com as outras e, portanto, as provas ou trimestres devem ser equalizados de alguma forma. Em alguns casos, essa equalização pode ser facilitada pelo desenho da pesquisa. No caso deste trabalho, por exemplo, as provas trimestrais puderam ser equalizadas por possuírem itens em comum. No caso dos trimestres, a estrutura rígida de ensino na terceira série (mesmos critérios de avaliação, mesmos tipos de aulas, etc.) possibilitou o tratamento de certas categorias como “itens semelhantes”. Mas isso pode não ocorrer em muitos casos.

4.2.2. Outras possibilidades

A multidimensionalidade do problema surgiu devido ao caráter longitudinal da análise. Em princípio, penso que esse não seria necessariamente um problema encontrado se o estudo da competência fosse de natureza transversal.

Outra possibilidade de tratamento é usar a nota como indicador de mudança na competência. Há pesquisas que indicam que as notas escolares têm alguma relação com diferenças de escore entre pré-testes e pós-testes (POPLUN, 2009). Dessa forma, poderia se encontrar alguma forma de relacionar a nota à variação da competência entre duas ocasiões, e não à competência em si.

4.2.3. Vantagens e desvantagens em relação a testes padronizados

Como já foi discutido nos primeiros capítulos, são vantagens do uso de avaliações escolares em relação ao uso de testes: (i) a sintonia das avaliações com o currículo real, (ii) o engajamento dos estudantes nas avaliações e (iii) a abundância de dados disponíveis e possibilidade de usar várias ondas de dados em um mesmo ano.

Essas vantagens podem ser reafirmadas após a conclusão de que o uso das avaliações é válido.

Há, entretanto, uma grande desvantagem: as restrições ao desenho metodológico impostas pela ética da prática educativa, que acabam levando às ressalvas mencionadas.

4.3. ALGUMAS PONDERAÇÕES

A limitação metodológica que pode ter levado ao decaimento da competência medida pela variável R(ITENS), já discutida em uma seção deste capítulo, me faz trazer uma questão para reflexão: se a competência em certo domínio (como a física) apresenta essa multidimensionalidade em relação aos subdomínios (mecânica, eletricidade,...), essa multidimensionalidade não pode também trazer problemas aos estudos longitudinais que usam testes padronizados?

De fato, essa questão já foi discutida por Reckase (2009). Ele sugere que os testes aplicados em diferentes ocasiões, mesmo se equalizados, podem refletir dimensões diferentes. Mesmo se o domínio específico não variar, o aumento da complexidade dos itens pode fazer com que sejam exigidos outros tipos habilidade para resolvê-los.

Além disso, como se espera que o aumento da competência de um grupo de estudantes ocorra de forma diferente para diferentes subdomínios, e que o maior crescimento ocorra no subdomínio que foi trabalhado em sala de aula (RECKASE, 2004; SAYRE e HECKLER, 2009), o não alinhamento dos testes com o currículo pode mascarar um aumento de competência (de forma semelhante, mas não igual, ao que ocorreu com nossa análise da variável R(ITENS)). E, ainda, o uso de um único conjunto de testes para comparar sujeitos

(ou escolas) submetidos a diferentes currículos, não pode medir precisamente o crescimento da competência de todos, pois o alinhamento dos conteúdos do teste com o currículo não ocorrerá em muitos casos.

De fato, parece que a questão da uni/multidimensionalidade é muito mais sutil e delicada do que este pesquisador supunha no início do estudo ou do que sugerem alguns estudos longitudinais de aprendizagem.

4.4. CONCLUSÔES

Nesta dissertação procurei investigar se avaliações escolares ordinárias podem ser usadas para estudar a evolução da competência em física. Para isso, utilizei dois tipos de avaliações: (i) notas trimestrais - uma avaliação global, envolvendo aspectos cognitivos, emotivos e motivacionais; e (ii) provas de itens dicotômicos – uma avaliação que, apesar de também envolver outros aspectos, está mais relacionada a aspectos cognitivos.

A partir das duas análises e da comparação entre elas, foram obtidas algumas conclusões consistentes. Dessa forma, defendo a idéia de que as avaliações escolares podem, sim, ser usadas no estudo de mudança, mas com algumas ressalvas. A primeira delas é que a competência medida depende do tipo de avaliação usada e, por isso, avaliações mais gerais, como a nota trimestral, ou o rendimento global (mais fáceis de obter em secretarias de escolas) levam a uma conceituação mais vaga da competência que está sendo medida. A segunda ressalva é que o contexto das avaliações deve possibilitar a equalização e a construção de uma escala estável para a competência. Uma última ressalva está em que um tratamento unidimensional pode levar a uma distorção na estimativa da variação da

competência com o tempo, dada a forma com que a avaliação escolar se relaciona com o conteúdo trabalhado em sala de aula.

A discussão do problema entre a relação do conteúdo trabalhado em sala com o conteúdo cobrado em avaliações me levou a questionar a possibilidade de um teste sistêmico único avaliar o crescimento da competência para sujeitos (ou escolas) com currículos diferentes.

4.5. LIMITAÇÔES DA PESQUISA E PESQUISAS FUTURAS

Apresento abaixo as três principais limitações desta pesquisa. As limitações, (i) e (ii), levaram à impossibilidade de um tratamento multidimensional. Pesquisas futuras sobre o uso de avaliações escolares ordinárias deveriam ser capazes de superar essas duas limitações (ou pelo menos uma delas) para que se possa ter uma melhor idéia de como se dá o crescimento da competência em física e de suas diversas componentes.

As principais limitações foram:

I- O uso de pouca variedade de avaliações

Isso me forçou a trabalhar com um conceito relativamente vago de competência. Outros tipos de avaliação (como respostas dissertativas, ou problemas abertos) poderiam possibilitar uma investigação mais detalhada de várias componentes da competência (ou várias competências) separadamente.

II- Restrição no desenho de "coleta" de dados

O fato de usar notas de avaliações fez com que não houvesse uma medida da competência nos subdomínios da física antes do período de instrução. Isso pode ter levado à impossibilidade de detectar o crescimento da competência nesses subdomínios e ao conseqüente viés negativo na inclinação da competência, conforme foi discutido.

III- Particularidade do contexto

A grande particularidade do contexto desta pesquisa – desde as características do alunado até a estrutura de curso de física da terceira série – gera uma enorme restrição para as possibilidades de generalização dos resultados.

Acho que esta dissertação pode ajudar na reflexão sobre o uso de avaliações escolares em pesquisas sobre competência. Em especial em estudos longitudinais de evolução da competência em física. Outros estudos podem vir a reforçar (ou a contradizer) os achados aqui expostos, para que se possa, em um futuro próximo, ter mais clareza sobre as possibilidades de uso dessa enorme quantidade de dados “coletados” regularmente em quase todas as escolas do Brasil e disponíveis nas secretarias das escolas.

CAPÍTULO 5: REFERÊNCIAS BIBLIOGRÁFICAS

1. AUSUBEL, D., NOVAK, J., & HANESIAN, H. *Educational Psychology:* A Cognitive View (2.ed.). Nova York: Holt, Rinehart & Winston, 1978.

2. AYALA, C. C.; SHAVELSON, R. J.; YIN, Y. Reasoning Dimensions Underlying Science Achievement: The Case of Performance Assessment. *Educational Assessment*. v. 8, n. 2, p. 101–121, 2002.

3. BAXTER, G. P.; GLASER, R. Investigating the Cognitive Complexity of Science Assessments. *Educational Measurement: Issues and Practice*, v.17, n.3, p.37-45, 1998.

4. BOND, G. T. e FOX, C. M. *Applying the Rasch Model*: Fundamental Measurement in the Human Sciences – 2.ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2007, 340p.

5. BROOKHART, S. M. Developing Measurement Theory for Classroom Assessment Purposes and Uses. *Educational Measurement: Issues and Practice*, v.22, n.4, p.5-12, 2003.

6. BYRNES, J. P.; MILLER, D. C. The Relative Importance of Predictors of Math and Science Achievement: An Opportunity-propensity analysis. Contemporary Educational Psychology, v.32, p.599-629, 2007.

7. CATSAMBIS, S. *Expanding Knowledge of Parental Involvement in Children's Secondary Education:* Connections with High School Senior's Academic Success. Social Psychology of Education, v.5, n.2, p.149-177, 2001.

8. COELHO, G. R.; BORGES, O. O Entendimento dos Estudantes Sobre a Natureza da Luz em um Currículo Recursivo. *Caderno Brasileiro de Ensino de Física*, v.27, n.1, p.63-87, 2010.

9. COHEN, P.; COHEN, J.; AIKEN, L. S.; WEST, S. G. The Problem of Units and the Circumstance for POMP. *Multivariate Behavioral Research*, v.34, n.3, p.315-346, 1999.

10. GOLDSTEIN, H.; HEALY, M. J. R. The Graphical Presentation of a Collection of Means. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, v.158, n.1, p.175-177, 1995.

11. GRIGG, W. S.; LAUKO, M. A.; BROCKWAY, D. M. The Nation's Report Card: Science 2005. 2006. Disponível em: <http://nces.ed.gov/nationsreportcard/science>.

12. JOARDER, A. H.; LATIF, R. M. Standard Deviation for Small Samples. *Teaching Statistics*, v.28, n.2, p.40-43, 2006.

13. JOHNSON, B. B. *An Examination of International Drivers of Educational Achievement*. 2009. Tese (Doutorado em Economia) – University of California, Berkeley, 2009.

14. JUSSIM, L. Grades May Reflect More Than Performance: Comment on Wentzel (1989). *Journal of Educational Psychology*, v.83, n.1, p.153-155, 1991.

15. KOEPPEN, K.; HARTIG, J.; ECKHARD, K.; LEUTNER, D. Current Issues in Competence Modeling and Assessment. *Zeitschrift für Psychologie / Journal of Psychology*, v.216, n.2, p.61–73, 2008.

16. LAWRENZ, F.; WOOD N. B.; KIRCHHOFF, A.; KIM, N. K.; EISENKRAFT, A. Variables affecting physics achievement. *Journal of Research in Science Teaching*, v.46, n.9, p.961-976, 2009.

17. LINACRE, J. M. (2010). Winsteps® (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.. Retrieved August 1, 2009. Disponível em <http://www.winsteps.com/>

18. MA, X.; WILKINS, J. L. M. The Development of Science Achievement in Middle and High Schoolr: Individual Differences and School Effects. *Evaluation Review*, v.26, n.4, p.395-417, 2002.

19. McMILLAN, J, H. Secondary Teachers' Classroom Assessment and Grading Practices. *Educational Measurement: Issues and Practice*. Vol.20, n.1, p.20-32, 2001.

20. McMILLAN, J. H. Understanding and Improving Teacher's Classroom Assessment Decision Making: Implications for Theory and Practice. *Educational Measurement: Issues and Practice*, v.22, n.4, p.34-43, 2003.

21. MEAD, R. *A Rasch Primer:* The Measurement Theory of Georg Rasch Psychometrics services research memorandum 2008-001. Maple Grove, MN: Data Recognition Corporation, 2008.

22. MULLER, P. A.; STAGE, F. K.; KINZIE, J. Science Achievement Growth Trajectories: Understanding Factors Related to Gender and Racial-Ethnic Differences in Precollege Science Achievement. *American Educational Research* Journal, v.38, n.4, p.981-1012, 2001.

23. NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS. (2005). The Nation's Report Card: Science. Retrieved August 8, 2008, Disponível em <http://nationsreportcard.gov/science_2005>

24. POMPLUN, M. R. Do Student Scores Measure Academic Growth? *Educational and Psychological Measurement*, v.69, n6, p.966-977, 2009.

25. RASCH, G. On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy* v.14, p.58-94, 1977.

26. RECKASE, M, D. The Real World is More Complicated than We Would Like. Journal of Educational and Behavioral Statistics, v.29, n.1, p.117-120, 2004.

27. RECKASE, M, D. Introdução. In:_____. *Multidimensional Item Response Theory.* Springer, Nova York, 2009, cap.1, p.1-10.

28. ROESER, R. W.; SHAVELSON, R. J.; KUPERMINTZ, H.; LAU, S.; AYALA, C.; HAYDEL, A.; SCHULTZ, S.; GALLAGHER, L.; QUIHUIS, G. The Concept of Aptitude and Multidimensional Validity Revisited. *Educational Assessment*,. v.8, n.2, p.191-205, 2002.

29. SAYRE, C. E.; HECKLER, A. F. Peaks and Decays of Student Knowledge in an Introductory E&M Course. *Physical Review Special Topics – Physics Education Research*, v. 5, 2009

30. SHAVELSON, R. J; ROESER, R. W.; KUPERMINTZ, H.; LAU, S.; AYALA, C.; HAYDEL, A.; SCHULTZ, S.; GALLAGHER, L.; QUIHUIS, G. Richard E. Snow's Remaking of the Concept of Aptitude and Multidimensional Test Validity: Introduction to the Special Issue. *Educational Assessment*, v.8, n.2, p.77-99, 2002.

31. SINGER, J. D.; WILLETT, J. B. *Applied Longitudinal Data Analysis*: Modeling Change and Event Occurrence. Nova York: Oxford University Press, 2003. 644p.

32. SIRIN, S. R. Socioeconomic Status and Academic Achievement: A Meta Analytic Review of Research. *Review of Educational Research*, v.75, n.3, p.417-453, 2005.

33. SMITH, J. K. Reconsidering Reliability in Classroom Assessment and Grading. *Educational Measurement: Issues and Practice*, v.22, n.4, p.26-33, 2003.

34. TRIOLA, M. F. *Introdução à Estatística*; 10. ed. Rio de Janeiro: LTC, 2008. 696p.

35. WENTZEL, K. R. Classroom Competence May Require More Than Intellectual Ability: Reply to Jussim (1991*). Journal of Educational Psychology*, v.83, n.1, p.156-158, 1991.

36. WHITE, K. R. The Relation between Socioeconomic Status and Academic Achievement. *Psychological Bulletin*, v.91, n.3, p.461-481, 1992.

37. WRIGHT, B. D. A History of Social Science Measurement. *Educational Measurement: Issues and Practice*, v.16, n.4, p.33-45, 1997.