MODELOS ESTATÍSTICOS E PONTOS DE CORTE PARA AVALIAÇÃO DE COMPETÊNCIAS COGNITIVAS E LINGUÍSTICAS DE IMIGRANTES.

STATISTICAL MODELS AND CUT-OFF SCORES FOR THE ASSESSMENT CONCERNING COGNITIVE AND LINGUISTIC SKILLS OF IMMIGRANTS.

Sandra Figueiredo[1]

PSIQUE • e-ISSN 2183-4806 • VOLUME XV • ISSUE FASCÍCULO 1
1st JANUARY JANEIRO - 30th JUNE JUNHO 2019 • PP .30-41
Submited on May 13th, 2019 | A ccepted on May 13th, 2019 (2 rounds of revision)
Submetido a 13 de maio, 2019 | A ceite a 22 de Março, 2019 (2 rondas de revisão)

Resumo

A literatura não apresenta instrumentos validados para medir diferentes componentes (linguísticos e cognitivos) de proficiência da população escolar imigrante, com entrada recente num país de acolhimento como Portugal. Por outro lado, os resultados de testes de avaliação de acuidade em Língua Segunda são difíceis de interpretar. Para examinar a validade de provas diagnósticas de proficiência em Língua Segunda, especificamente para as dimensões linguística (ortográfica, semântica, sintática, lexical) e cognitiva (raciocínio verbal, decisão lexical, ritmo prosódico), é necessário determinar pontos de corte controlando as seguintes variáveis: idade cronológica (7-17 anos), idade de imigração, nacionalidade da criança, nacionalidade dos pais, diversidade das línguas maternas e residência no país de acolhimento num período < oito anos. No estudo que se encontra em curso, a análise estatística será feita com recurso à análise discriminante e serão medidas a sensibilidade e a especificidade (comparação entre os acertos e erros) através do modelo Receiver Operating Characteristics (ROC). É esperado que os grupos se diferenciem nas provas de acordo com as variáveis selecionadas e essa diferenciação sustente a definição de pontos de corte distintos considerando o nível de dificuldade dos sujeitos, o tipo de provas da bateria e a idiossincrasia do nosso contexto de acolhimento escolar.

Palavras-chaves: Língua segunda, Imigração escolar, Pontos de corte, Modelos estatísticos, Validação.

[1] Professora Associada do Departamento de Psicologia da Universidade Autónoma de Lisboa Luís de Camões; Investigadora Integrada da Fundação para a Ciência e Tecnologia (FCT) e do Centro de Investigação em Psicologia (CIP), Departamento de Psicologia da Universidade Autónoma de Lisboa Luís de Camões; Investigadora Colaboradora do Centro de Investigação em Educação (CIE) do ISPA – Instituto Universitário. E-mail: sfigueiredo@autonoma.pt.

Abstract

The literature shows no validated instruments for measuring various components (cognitive and language) of proficiency of immigrant school population, with recent entry in a host country such as Portugal. On the other hand, the results of tests for the evaluation of Second language acuity are difficult to interpret. To examine the validity of diagnostic evidence of second-language proficiency, specifically for the linguistic dimensions (spelling, semantic, syntactic, lexical) and cognitive (verbal reasoning, lexical decision, rhythm prosodic), it is necessary determine cut-off points controlling the following variables: chronological age (7-17 years), age of immigration, nationality of the child, the parents ' nationality, diversity of mother tongues and residence in the host country within < eight years. In the study that is underway, the statistical analysis will be made using discriminant analysis and shall be measured sensitivity and specificity (comparison between the hits and errors) through the model Receiver Operating Characteristics (ROC). It is expected that the groups differ in accordance with the selected variables and this distinction holds up the definition of distinct cut-off points whereas the level of difficulty of the subject, the type of evidence and the idiosyncrasies of our host school context.

Key words: Second language, Immigration, Cut-off points, Statistical models, Validation.

Avaliação linguística e cognitiva

A investigação sobre a avaliação de competências de indivíduos imigrantes, sobretudo imigrantes de segunda geração, tem assumido três perspetivas. Primeiro, a análise sobre a forma como os testes de avaliação em Língua Segunda (L2), nos países de acolhimento, têm sido utilizados em específicas populações não nativas para avaliar níveis da proficiência. Segundo, os testes têm verificado o efeito moderador de fatores para explicar a variabilidade da competência e do desempenho dessas populações (Cummins, 2008). Terceiro, os testes de diagnóstico de proficiência têm sido utilizados em estudos comparativos com outro tipo de testes (provas de fluência, provas de medição de controlo executivo, de tempo de reação, de audição dicótica) para comparar desempenhos em crianças e adultos (Bialystok & Feng, 2009; Hull & Vaid, 2007).

São objetivos de medida distintos, mas que são, frequentemente, confundidos por examinadores e por investigadores com a avaliação de proficiência. De facto, a primeira perspetiva pretende medir proficiência (Cutting & Scarborough, 2006), a segunda pretende identificar preditores ou moderadores de proficiência (Portes & Hao, 2002), a terceira ocupa-se da comparação de resultados entre testes de origem distinta, na educação e na psicologia (Bialystok & Feng, 2009; Derwing, Rossiter, Munro & Thomson, 2004). Os testes que têm sido desenvolvidos, explorados e validados na literatura internacional, visam populações escolares e não escolares imigrantes. No que respeita a populações escolares, a proficiência pode ser medida por componentes ou modalidades (Sawaki, 2007; Llosa, 2008). Por um lado, o vocabulário, semântica, sintaxe, ortografia, por outro lado, o raciocínio verbal, a memória, a atenção, a decisão lexical, a modificação morfológica, a compreensão metafórica, o ritmo prosódico. Considerando a diversidade de componentes, diferentes testes devem ser adequados para cada um desses componentes (Sawaki, 2007).

Desta forma, diferentes itens devem ser considerados para avaliação de cada uma das dimensões da proficiência, distinguindo ainda competência cognitiva de competência linguística. A utilização de itens de analogias verbais, por exemplo, avalia não só aspetos da linguagem, como também aspetos cognitivos. A utilização de itens de nomeação verbal serve, sobretudo, a avaliação linguística. Para alguns autores, esta divisão não pode ser claramente separada pois o estímulo apresentado –visual ou auditivo – e a técnica – eye tracking, lista de palavras em formato papel e lápis, descodificação com ficheiros áudio – varia

o foco da medida do teste e pode abranger capacidades de ordem não apenas linguística, mas também cognitiva (Bialystok & Feng, 2009; Blumenfeld & Marian, 2007; Song, 2008; Vandergrift & Goh, 2012). Para outros autores, testes como palavras cognatas estão fortemente correlacionados, em medida linguística e cognitiva, com outros testes como o de vocabulário o que dificulta a escala de pontos de corte ("overlap" e "halo effect", Sawaki, 2007, p. 381). Igualmente, Sawaki (2007) sugere que os investigadores e examinadores possam testar a individualidade de cada medida (e cada pontuação a obter) em períodos de tempo (de aplicação) diferentes para evitar a análise de "overlap".

Para uma avaliação compreensiva, e considerando a definição e validação de itens num teste, então deve ser desenvolvida uma bateria de tarefas que permita captar o maior número possível de indicadores da proficiência. A proficiência tem sido bastante examinada e validada no domínio funcional, ou seja, na determinação de níveis relacionados com as competências comunicativas (American Council on the Teaching of Foreign Languages; Comissão Europeia, 2001) e nem sempre de forma consonante entre a avaliação de scores mínimos de proficiência e a avaliação de competências mínimas para entrar no contexto académico de acolhimento (Feast, 2002; Harsch & Rupp, 2011). Um dos principais problemas é a parca validação desses níveis atribuídos e que se designam de níveis de proficiência, mas que na verdade são descritores de metas de proficiência (Fulcher, 2010; Sawaki, 2007; Harsh & Rupp, 2015; Tannenbaum & Cho, 2014). Os descritores são importantes para a orientação da comunicação interpessoal do indivíduo, mas não servem para explicar as diferenciações linguísticas e cognitivas de indivíduos em contexto de aprendizagem e de avaliação (Feast, 2002; Harsh & Rupp, 2011). Também não respondem satisfatoriamente à análise dos preditores envolvidos na diferenciação entre grupos de aprendentes de L2 e quanto aos níveis mais preocupantes, B1 e B2 (Harsch & Rupp, 2011), sendo esses níveis os que mais identificam os jovens imigrantes após um período breve de exposição à L2.

Ainda sobre os descritores e as normas de orientação, aplicando-se ao cenário europeu, a EALTA tem desenvolvido um importante trabalho no sentido de definir orientações para as provas a aplicar e as respetivas especificidades com o objetivo de evitar a marginalização dos alunos não nativos em espaços europeus (Kaftandjieva, 2010). Notavelmente há algumas diretrizes no sentido de esclarecer as normas dos testes a aplicar (Kaftandjieva, 2010), mas, de novo, sem indicação de pontos de corte para essas provas. Entendemos, então, que primeiro, antes do objetivo da criação de itens para testes a validar em populações não nativas, em condição de exame e de aprendizagem de Língua Segunda, a investigação deve explorar as diferenças e a "reclassificação" de desempenho entre minorias imigrantes e étnicas no que respeita ao seu comportamento verbal em específicos itens de medição linguística e cognitiva (Dornyei, 2014; Kaftandjieva, 2010; Mahoney, Haladyna & MacSwan, 2009). Grupos de nacionalidade com origem em países com sistemas educativos diferenciados em termos de recursos e de preparação pedagógica apresentam um background que produz um efeito moderador muito significativo nos resultados escolares das recentes gerações de imigrantes (Magnuson, Lahaie & Waldfogel, 2006; Oh & Fuligni, 2010). Por outro lado, o treino dos examinadores para a aplicação correta dos itens é fundamental para que os pontos de corte de testes funcionem para estas populações (Harsch & Rupp, 2011).

Também o período em que se aplicam esses testes, utilizando baterias mais ou menos compreensivas (em termos de número de itens), é uma variável explicativa dessas diferenças: o tempo de residência no país de acolhimento e a idade cronológica do indivíduo. Outros fatores deveriam ser explorados no estudo dos preditores da proficiência e de aptidão para a aprendizagem em países de acolhimento tais como a hora do dia em que se aplicam os itens de testes de avaliação da proficiência linguística e cognitiva. Mas desconhecem-se estudos nesta área específica. Sendo duas competências diferentes, a linguística e a cognitiva, a predisposição do aprendente, jovem ou adulto, para responder com melhor desempenho pode variar de acordo com as suas preferências cognitivas previamente aprendidas, preferências determinadas biologicamente pelo tipo diurno (Goldstein, Hahn, Hasher, Wiprzycka & Zelazo, 2007).

Segundo, com o conhecimento dos preditores de competências linguísticas e cognitivas de aprendentes não nativos, os itens são testados para a composição de uma bateria válida de avaliação que possa ser utilizada em contextos sobretudo escolares (Portes & Hao, 2002). A preocupação maior é a aplicação

de testes previamente validados para diferentes populações imigrantes e étnicas no período de chegada ao país de destino (Van Tubergen & Kalmijn, 2005). Estudos revelam que as populações com experiência recente de imigração apresentam-se em condição fragilizada (em período crítico para a aprendizagem e adaptação) que pode comprometer a sua evolução académica e profissional (Berry, Phinney, Sam & Vedder, 2006; Pereira & Ornelas, 2011).

Esse período crítico é comprometido quando não se identificam as reais lacunas de competência através de testes validados e diferenciados de acordo com as minorias de que fazem parte crianças e jovens (Pumariega, Rothe & Pumariega, 2005). Nesta identificação, que se relaciona com o necessário estudo prévio sobre preditores para o exame de populações imigrantes, a nacionalidade dos pais e o investimento parental é fundamental (Altschul,2011; Lahaie, 2008; Ong, Phinney & Dennis, 2006), com evidência do papel da mãe imigrante para o desempenho dos filhos (não nascidos no país de destino) (Sohn & Wang, 2006).

Modelos estatísticos e análise de pontos de corte

Em contexto de teste diagnóstico e de avaliação periódica de aprendizagem de alunos imigrantes, considerando um segmento alargado de idades, é imprescindível o conhecimento de procedimentos de adaptação e de cálculo de pontuação para a definição de pontos de corte. Os resultados de testes de diagnóstico de competências e de desempenho aplicados a alunos imigrantes são de difícil interpretação para professores e educadores, mas também para os próprios alunos (Chua, Liow & Yeong, 2016). São os pontos de corte que permitirão a investigadores e profissionais de educação compreender que estão a avaliar componentes específicos e com valores segundo uma norma previamente validada.

No entanto há pouca evidência disponível sobre o racional dos pontos de corte para este tipo de testes e para esta população como anteriormente referido (Kaftandjieva, 2010). A literatura internacional tem desenvolvido bons progressos para a definição de níveis qualitativos de proficiência, mas apenas no domínio linguístico e funcional (American Council on the Teaching of Foreign Languages; Comissão Europeia, 2001). Atendendo às diferenças de desempenho que grupos de imigrantes apresentam em diferentes tipos de tarefas, provavelmente os pontos de corte devem ser baseados por critérios relacionados com as diferenças sociodemográficas dos sujeitos. Em estudo prévio (Figueiredo & Silva, 2009; Figueiredo 2010) com amostra de casos (imigrantes em Portugal) e amostra de controlo (nativos, portugueses) definimos como ponto de corte o resultado de uma forma de cálculo que se baseou na pontuação total (nota total de nove testes) dos testes resolvidos. No entanto, o cálculo teve em conta as diferenças a considerar na população da amostra, especificamente no que respeita à idade e ao ano de escolaridade. O cálculo foi considerado da seguinte forma (Figueiredo, 2010, p. 349):

Fórmula: PC = [(M casos * DP casos) + M controlos * DP controlos)] / (DP casos + DP controlos)

$$\mathbf{PC} = \frac{Mcasos * DPcasos + Mcontrolos * DPcontrolos}{DPcasos + DPcontrolos}$$

Por um lado, não se encontram estudos que comparem testes e pontos de corte obtidos para os mesmos grupos ou minorias linguísticas e étnicas (sobre o conceito de 'imigrante', 'minoria', 'grupo 'étnico', ver Figueiredo, 2017 e American Psychological Association [APA], 2000) em contextos de acolhimento (por 'contextos' entenda-se língua-alvo e país de destino), como os há para outras áreas (como na área da saúde ou da avaliação psicológica) quanto a comparações de testes e pontos de corte em populações de diferentes países (Saxena, Ambler, Cole & Majeed, 2004; Wang & Wang, 2002). Por outro lado, verificou-se estudos muito particulares sobre os pontos de corte para populações imigrantes na circunstância de aprendentes

de L2, mas no domínio da "Fundamental Difference Hypothesis" de Bley-Vroman (2018) e focando tarefas com proeminência da componente gramatical (DeKeyser, 2000; Sawaki, 2007).

Kaftandjieva (2010) analisou os métodos existentes para a definição de testes e dos seus pontos de corte ou formatos, focando contextos europeus. A fragilidade de testes de desempenho, no que respeita à questão de definição de pontos de corte válidos, começa nos seus formatos simples de escolha múltipla ("centered-tests"), pois são estes que geram maior ambiguidade. Os examinadores atribuem valores subjetivos aos resultados a partir desses testes (p. 33). Pelo contrário, os testes de desempenho que se designam como "performance-centered" (pp. 33-34), ou seja, que focam o tipo de indivíduos testados, são os que menos problema geram para a definição de um ponto de corte e que registam o desempenho e as caraterísticas do sujeito. Os autores concluem que os melhores testes, para melhores pontos de corte ou normas, serão os "performance-centered", porém são menos observados porque demandam uma base de estudo empírico, um estudo piloto prévio e um específico período de tempo para serem testados. É neste âmbito que a investigação em L2 deve responder com rigor, seja com estudos transversais, seja com longitudinais. O estudo de Kaftandjieva (2010) permite verificarmos também que a tentativa não eficaz de estabelecer pontos de corte tem sido explicada pelas análises estatísticas incipientes que só recorrem a análises descritivas, ou só a análises regressivas ou só a análises de "item response theory". Mais, os autores igualmente valorizam a existência de mais do que um ponto de corte para os testes, considerando a variação interindividual.

Nesta linha de ideias, a medida de sensibilidade e especificidade (calculadas pelo modelo receiver operator characteristics - ROC) apresenta-se como a opção mais completa e válida para determinar o ponto de corte (ou pontos de corte) de testes. De acordo com Koen, Barrett, Harlow e Yonelinas (2017), podemos obter resultados de confiança e estimativa por máxima verosimilhança de forma a permitir aos investigadores implementar de forma mais facilitada os específicos instrumentos (compreender os pontos de corte) para as mesmas populações e metas (como o caso de testes de avaliação linguística e cognitiva). E, também, visando o "end-user" (p. 1400) como referem Koen et al. (2017), sendo no nosso caso o "end-user" identificado como: professores, educadores, psicólogos e os próprios examinandos. Este método de análise estatística (ROC) possibilita o entendimento, de forma mais consensual, acerca dos processos inerentes ao desempenho dos sujeitos examinados em determinados testes.

Baterias de testes compreensivas e a teoria do fator G

Considerando o número incipiente de medidas validadas para a avaliação linguística e para o raciocínio verbal, sobretudo considerando o contexto europeu e as suas novas populações não nativas, a criação de uma bateria de testes será importante para designar não apenas um ponto de corte (de toda a bateria) mas vários para cada uma das tarefas que completam a referida bateria. Assim apresenta-se um modelo de avaliação compreensiva para, através de várias provas num só teste, podermos diminuir o erro de teste ao atender às diferenças de cada indivíduo (identificado num perfil que o ponto de corte, de cada tarefa, providencia de acordo com o grupo ou minoria em que se integra) e à variabilidade esperada em diferentes provas pois avaliam diferentes componentes como anteriormente referido. Atente-se que em estudo prévio (Figueiredo, 2010) já explorámos a nota total e ponto de corte considerando a perspetiva geral de uma pontuação por grupos etários, sobretudo. Consideramos que daqui deveremos partir para pontos de corte e normas diferenciadas por teste e grupos, portanto.

Sobre a variabilidade, a relevância de testes e a avaliação de facetas (itens, tipos de score atribuídos, perfil de quem avalia) dos testes para população imigrante, desde cedo foram desenvolvidos estudos na área da "Generalizability theory" (Brennan, 1992; Lynch & McNamara, 1998) ou teoria dos testes de fator G (Almeida & Lemos, 2005; Primi, 2003) para permitir compreender a validade e a decisão sobre a relevância de itens por parte dos investigadores. Este tipo de estudos na área da teoria do fator G aproxima-se mais da intenção de analisar a diferenciação de itens e da necessária relativização de scores para este tipo de populações em avaliação. Poderá ser determinado um critério ou um score geral (Lynch & McNamara, 1998) para cada teste, mas respeitando as especificidades do sujeito avaliado (Dornyei, 2014) e das chamadas 'facetas'

que interferem necessaria e naturalmente no processo de teste (Bachman, Lynch & Mason, 1995). Uma delas deve ser identificada precisamente no perfil do grupo ou minoria, portanto considerando o sujeito avaliado.

Mais, é importante atentar na probabilidade de nem todos os testes ou tarefas de uma bateria terem de ser aplicados, em conjunto, a um grupo de crianças (ou grupo de adultos) em contexto escolar o que justifica o princípio da variação e da "dependability" (Bachman, Lynch & Mason, 1995; Lynch & McNamara, 1998). Há testes que poderão não ser adequados para determinadas minorias linguísticas ou étnicas porque não conferem importância à avaliação de que se necessita para tais grupos. Essa probabilidade irá afetar a validade do ponto de corte geral da bateria se este apenas for utilizado como única opção (um ponto de corte, da bateria total) apenas pelos investigadores, professores ou administradores. Daí a importância de assumir a relativização dos scores porque há caraterísticas que variam entre os indivíduos pela sua condição. Verificou-se, por exemplo, num estudo preliminar (Figueiredo, 2010), que testes com estímulo auditivo são distintivos de grupos de crianças imigrantes, na escola, mas não servem para diferenciar, numa fase inicial que coincida com a recente entrada em Portugal (< 1 ano), crianças de específicas minorias tais como as que têm origem nos países do subcontinente indiano e nos países da Europa de Leste.

Ainda sobre os estudos da teoria G, Lynch e McNamara (1998) desenvolveram uma investigação, com duas décadas, sobre proficiência em Inglês, na Austrália, em que na combinação de fatores de análise para determinação de scores de itens, verificaram que o fator que mais explicou a variância no desempenho dos testes foi o 'sujeito testado' comparativamente com os outros fatores tais como o avaliador e a dificuldade do item ou tipo de teste (p. 166). Diferentes scores foram também verificados pela 'severidade' distinta dos examinadores. Num estudo mais recente (Harsch & Rupp, 2011) o fator 'sujeito testado' também foi o que mais se destacou para a tentativa de determinação de pontos de corte em acordo com os níveis sugeridos pelo QECR (2001; 2003). Um dos aspetos que pode auxiliar a determinação de scores e de custos para a comunidade escolar e académica será também o fator 'tempo'. A duração dos testes é recomendada como menor para um desempenho mais apurado. Sobre este fator e outro – a influência do interlocutor nos testes específicos de oralidade e compreensão oral – Fulcher (2014) refere a mesma explicação para a variabilidade dos scores dos sujeitos em condição de L2.

Tratando-se de um tipo de testes de avaliação de componentes de proficiência em L2, para assegurar o sucesso e a integração escolar de populações não sinalizadas com perturbações específicas de linguagem ou outras patologias, dever-se-ia analisar se esses testes devem ser orientados apenas na L2 ou simultaneamente na L2 e na Língua Materna (L1). Apesar da importância de competências consolidadas na L1 para a aprendizagem de novas línguas (Vandergrift & Baker, 2015), não se conhece bem a relação entre a medição de limitações em componentes de L2 e a medição de componentes equivalentes em L1 (Ellis, 2004; Nassaji, 2003). Mesmo os especialistas na aplicação de testes psicoeducacionais, considerando populações imigrantes com perturbações da linguagem, não se encontram familiarizados com as propriedades de testes em L2, menos ainda com a compreensão das especificidades das minorias (APA, 2000).

A dificuldade em compreender os testes de L2, por profissionais e investigadores é acrescida pela sobrevalorização do efeito de específicas variáveis tais como o nível socioconómico (SES) dos alunos imigrantes que está bastante analisado como um preditor das competências em L2 e muito associado aos grupos hispânicos no contexto norte-americano como país de acolhimento (Dockrell, Stuart, & King, 2004; Mollborn, Fomby, & Dennis, 2012). Esta correlação não é necessariamente aplicável a populações imigrantes noutros países recentemente reconhecidos como país de acolhimento como o caso de Portugal (Figueiredo, 2017). Pelo contrário, outros grupos que se comportam de forma exímia em escolas norte-americanas, como os locutores indo-iranianos, têm desempenhos muito fracos em testes no contexto de escolas portuguesas (Figueiredo, 2017). E o SES é um fator paralelo a outros fatores tais como a idade, a nacionalidade da mãe, o tipo de escola frequentada e, sobretudo, o tipo de recursos e testes aplicados nessas escolas.

Esses testes variam entre escolas com implicações para os resultados globais que são de conhecimento público (OECD, 2016; Marôco, Gonçalves, Lourenço & Mendes, 2018). As caraterísticas psicológicas

tais como os estilos de aprendizagem e as atitudes dos alunos são variáveis distintivas no processo de teste de competências (Dornyei, 2014). Essas implicações não são, porém, necessariamente positivas pois muitos dos testes aplicados, mesmo no contexto da avaliação do Programa Internacional de Avaliação de Alunos (PISA), só se referem a itens específicos e não a uma avaliação puramente compreensiva – "single item assessment deficit" (Anderson, Mak, Chahi & Bialystok, 2018). Ainda, de acordo com o grupo ou minoria linguística a que se refira o estudo de proficiência e os estudos comparativos, os preditores como idade, data de chegada e nacionalidade variam de forma significativa quanto ao seu efeito no desempenho dos grupos (Guven & Islam, 2015).

Também os indivíduos imigrantes inseridos em turmas com nativos e outros pares imigrantes com elevados níveis de desempenho pode ser um stressor que dificulta que os primeiros evidenciem um desempenho equivalente à real competência em L2 e, por consequente, nas várias unidades curriculares do programa escolar. Por isto, os testes, com pontos de corte validados para diferentes nacionalidades, devem ser aplicados no período de chegada à escola e, depois, periodicamente (fim dos períodos letivos e fim do ano letivo). Todavia, e como já referido, a literatura dificilmente apresenta estudos sobre a definição de pontos de corte para itens de testes para avaliação linguística, mas também para a avaliação cognitiva no que concerne a estas populações. Ou só reporta aos contextos em Inglês como L2 desde os primeiros estudos de validade para testes e investigadores (Lynch & MacNamara, 1998).

Razoabilidade e ambiguidade das provas de avaliação

Itens de testes como o teste de palavras cognatas (Malabonga, Kenyon, Carlo, August & Louguit, 2008) têm sido bastante aplicados em populações imigrantes (Scarcella & Zimmerman, 2005; Sawaki, 2007), portanto para a avaliação escolar em L2, mas com pontos de corte validados sobretudo para populações hispânicas em contexto de Inglês como L2 (Malabonga et al., 2008). Por outro lado, itens de nomeação de imagens, de compreensão auditiva e de analogias verbais do Woodcock Language Proficiency Battery-Revised (WLPB-R, 1991) são bastante utilizados em populações similares. E os testes de convergência para validação dos itens utilizam estes mesmos conjuntos de testes que normalmente têm aferição principal com amostras hispânicas. Outros estudos de validação para testes como o "extract the base test" (Goodwin, Huggins, Carlo, Malabonga, Kenyon, Louguit et al., 2012) foram desenvolvidos no contexto do Inglês como L2 e com as minorias imigrantes norte-americanas mais representativas. A maioria destes itens é testada quanto à sua correlação com medidas de identificação lexical, compreensão escrita e outras medidas de decisão lexical (Goodwin et al., 2012).

Na última década tem sido revista a validade dos testes considerando a ambiguidade de procedimentos e de pontos de corte para aprendentes de L2 (Mahoney & MacSwan, 2005). Na esfera da ambiguidade de procedimentos e das limitações dos testes por serem maioritariamente validados em contextos norte-americanos deve ser considerada a avaliação da razoabilidade de dificuldade de itens ("Differential item functioning") dos testes linguísticos para a segunda geração de imigrantes (Uiterwijk & Vallen, 2005) e, para essa dificuldade dos itens, ter em conta a influência do background linguístico dos examinandos (Shin, 2005). O mesmo item ou teste, pressuposto para medir uma determinada modalidade linguística ou cognitiva, pode ser diferentemente respondido por indivíduos imigrantes dada a sua compreensão limitada ou diferente face ao input linguístico (o enunciado conforme apresentado para cada teste). Isto acontece igualmente para os casos da dificuldade e da ambiguidade de itens em testes de matemática que são menos válidos porque têm limitações linguísticas (para a segunda geração de aprendentes de L2) nos enunciados (Uiterwijk & Vallen, 2005).

Também cada item, na avaliação de segunda geração de imigrantes, pode estar a avaliar mais do que o que realmente se propõe medir (Uiterwijk & Vallen, 2005). No entanto, consideramos que pode isso não ser perturbador se o examinador perceber o procedimento e o constructo dos testes. Conforme referido anteriormente, uma tarefa de palavras cognatas mede aspetos linguísticos, mas também cognitivos (raciocínio verbal, fluência semântica, transferência entre L1 e L2). Os testes de compreensão e de produção

oral são os tipos de itens mais aplicados no contexto de avaliação de L2 (Bernstein, Van Moere & Cheng, 2010). Para a questão central da determinação de pontos de corte, Fulcher (2014) refere a importância da interferência do interlocutor (de quem apresenta o teste, não só quem o avalia) pois o input apresentado ao sujeito em período de teste pode gerar distratores e ruído na avaliação. A avaliação da compreensão e da produção oral tem sido o principal foco dos testes internacionais de proficiência (Feast, 2002), mas com pontos de corte testados de forma insuficiente. Pelo contrário, testes semânticos (por exemplo, testes com similaridades/sinónimos) são mais recentes e têm produzido evidência sobre a necessidade de gerar pontos de corte dadas as diferenças observadas entre alunos com backgrounds linguísticos diferentes (Pena, Bedore & Rappazzo, 2003).

No caso dos testes de vocabulário, estudos asseveram a necessidade de atualizar pontos de corte ou "mínimos" expectáveis para o domínio lexical dos indivíduos nãos nativos em contexto escolar (Hazenberg & Hulstun, 1996). Neste estudo, na década de noventa, Hazenberg e Hulstun (1996) validaram uma nova referência como mínimo de frequência lexical para alunos universitários (no primeiro ano do Ensino Superior) terem sucesso académico. Também examinando a mesma frequência lexical necessária em amostras nativas para o mesmo objetivo. Concluíram o mínimo de 10 000 palavras como domínio lexical para os alunos não nativos (>18 anos) por contraste com o mínimo estipulado de 3000 a 5000 palavras em estudos prévios similares. Mais recentemente Schmitt e Schmitt (2014) verificaram novos pontos de corte, próximos aos anteriores, mas com refinamento em termos de níveis e não apenas considerando uma norma de frequência lexical. Atualmente é importante replicar o princípio de estudos como estes para observar como a populações académicas e imigrantes necessitam de ser avaliadas, de forma válida e adaptada ao seu contexto (país de acolhimento, reportando aos imigrantes).

Em suma, pretende-se determinar pontos de corte para cada prova, atendendo às diferenças de idade, género, nacionalidade, idade de aquisição da L2, entre outros fatores, com o objetivo de providenciar aos profissionais de educação e aos estudantes uma norma diagnóstica válida e mais congruente em Portugal. Este estudo em desenvolvimento, assente na literatura anteriormente revista e sobretudo focando os estudos de validação internacionais e as peculiaridades do desempenho das minorias imigrantes, avaliará a acuidade e a validade de provas diagnósticas de proficiência em Português como Língua Segunda, especificamente para as dimensões linguística (ortográfica, semântica, sintática, lexical) e cognitiva (raciocínio verbal, decisão lexical, ritmo prosódico). Desta forma, após o estudo empírico com população imigrante escolar portuguesa, está a ser concluída a fase de análise de pontos de corte com base nos resultados de desempenho obtidos para cada tarefa respondida por grupos diferenciados imigrantes. O cálculo de pontuação terá em conta as variáveis idade cronológica, idade de imigração, tipo de língua materna, hora do dia do preenchimento dos testes, nacionalidade da criança e nacionalidade dos pais.

Referências

Almeida, L. S., & Lemos, G. C. (2005). *Aptidões cognitivas e rendimento académico: A validade preditiva dos testes de inteligência*. Repositório da Universidade de Évora. Disponível em: http://dspace.uevora.pt/rdpc/handle/10174/1821

American Psychological Association (APA). (2000). *Guidelines for research in ethnic minority communities;* APA: Washington, DC, USA.

American Council on the Teaching of Foreign Languages, https://www.actfl.org/

Anderson, J. A., Mak, L., Chahi, A. K., & Bialystok, E. (2018). The language and social background question naire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods*, 50(1), 250-263. doi: 10.3758/s13428-017-0867-9

Altschul, I. (2011). Parental involvement and the academic achievement of Mexican American youths: what kinds of involvement in youths' education matter most? *Social Work Research*, 35(3), 159-170. doi: 10.1093/swr/35.3.159.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12(2), 238-257.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355-377. doi: 10.1177/0265532210364404.

Berry, J. W., Phinney, J. S., Sam, D. L., & Vedder, P. E. (Eds.) (2006). *Immigrant youth in cultural transition: Acculturation, identity, and adaptation across national contexts*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Bialystok, E., & Feng, X. (2009). Language proficiency and executive control in proactive interference: Evidence from monolingual and bilingual children and adults. *Brain and language*, 109(2-3), 93-100. Doi: 10.1016/j.bandl.2008.09.001.

Bley-Vroman, R. (2018). Language as "something strange". *Bilingualism: Language and Cognition*, 21(5), 913-914.

Blumenfeld, H. K., & Marian, V. (2007). Constraints on parallel activation in bilingual spoken language processing: Examining proficiency and lexical status using eye-tracking. *Language and cognitive processes*, 22(5), 633-660. doi: 10.1080/01690960601000746.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice,* 11(4), 27-34.

Chua, S. M., Rickard Liow, S. J., & Yeong, S. H. (2016). Using spelling to screen bilingual kindergarteners at risk for reading difficulties. *Journal of learning disabilities*, 49(3), 227-239.

Comissão Europeia (2001). *Quadro europeu comum de referência para as línguas: Aprendizagem, ensino, avaliação*. Porto: Asa.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In *Encyclopedia of language and education* (pp. 487-499). Boston: Springer.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific studies of reading*, 10(3), 277-299. doi: 10.1007/s11881-009-0022-0.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in second language acquisition,* 22(4), 499-533. doi:10.13140/RG.2.1.3958.4486.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language learning,* 54(4), 655-679. doi: 10.1111/j.1467-9922.2004.00282.x.

Dockrell, J., Stuart, M., & King, D. (2004). Supporting early oral language. *Literacy Today*, 40, 16–17. doi: 10.1348/000709910X493080.

Dörnyei, Z. (2014). *The psychology of the language learner: Individual differences in second language acquisition*. Routledge.

Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language learning,* 54(2), 227-275.doi: 10.1111/j.1467-9922.2004.00255.x.

Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal,* 3(4),70-85.Disponível em: http://ehlt.flinders.edu.au/education/iej/articles/v3n4/feast/paper.pdf

Figueiredo, S., & Silva, C. (2009). Decoding behaviour and connectivity mental system in second language context: critical period and the lateralization of language function. In K. Fanti (Ed.). *Applying Psychological Research to Understand and Promote the Well-being of Clinical and Non-clinical Populations* (pp. 29-41). Atenas: Athens Institute for education and Research.

Figueiredo, S. (2010). *Factores psicológicos e desempenho cognitivo na aprendizagem linguística.* (Tese de Doutoramento não publicada). Universidade de Aveiro, Aveiro. Portugal. Disponível em: http://opac.ua.pt/F/T6M7E2YQ2RDVSCFJSG8PEA71LQ1X4B8CY2MGCJ131F5T1MTSPM-23591?func=full-set-set&set_number=004886&set_entry=000001&format=999

Figueiredo, S. (2017). *Learning Portuguese as a Second Language. Boston: Springer International Publishing.* doi: 10.1007/978-3-319-55819-6

Fulcher, G. (2010). The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. In *A.Psyaltou-Joycey, & M. Matthaioudakis (Eds.) Advances in research on language acquisition and Teaching: Selected Papers* (pp. 15-26). Thessaloniki: GALA.

Fulcher, G. (2014). *Testing second language speaking.* Routledge.

Goodwin, A. P., Huggins, A. C., Carlo, M., Malabonga, V., Kenyon, D., Louguit, M., & August, D. (2012). Development and validation of extract the base: an English derivational morphology test for third through fifth grade monolingual students and Spanish-speaking English language learners. *Language Testing,* 29(2), 265-289.

Goldstein, D., Hahn, C. S., Hasher, L., Wiprzycka, U. J., & Zelazo, P. D. (2007). Time of day, intellectual performance, and behavioral problems in morning versus evening type adolescents: Is there a synchrony effect? *Personality and individual Differences*, 42(3), 431-440. doi: 10.1016/j.paid.2006.07.008.

Guven, C., & Islam, A. (2015). Age at migration, language proficiency, and socioeconomic outcomes: evidence from Australia. *Demography,* 52(2), 513-542. doi: 10.1007/s13524-015-0373-6

Harsch, C., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, 8(1), 1-33.

Hazenberg, S., & Hulstun, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied linguistics,* 17(2), 145-163. doi: 10.1093/applin/17.2.145

Hull, R., & Vaid, J. (2007). Bilingual language lateralization: A meta-analytic tale of two hemispheres. *Neuropsychologia,* 45(9), 1987-2008. doi: 10.1016/j.neuropsychologia.2007.03.002.

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests.* Cito, Arnhem: EALTA. Disponível em: http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf

Koen, J. D., Barrett, F. S., Harlow, I. M., & Yonelinas, A. P. (2017). The ROC Toolbox: A toolbox for analyzing receiver-operating characteristics derived from confidence ratings. *Behavior research methods,* 49(4), 1399-1406. doi: 10.3758/s13428-016-0796-z.

Lahaie, C. (2008). School readiness of children of immigrants: Does parental involvement play a role? *Social Science Quarterly,* 89(3), 684-705. Doi: 10.1111/j.1540-6237.2008.00554.x.

Llosa, L. (2008). Building and Supporting a Validity Argument for a Standards Based Classroom Assessment of English Proficiency Based on Teacher Judgments. *Educational Measurement: Issues and Practice,* 27(3), 32-42. Doi: 10.1111/j.1745-3992.2008.00126.x

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing,* 15(2), 158-180.

Magnuson, K., Lahaie, C., & Waldfogel, J. (2006). Preschool and school readiness of children of immigrants. *Social science quarterly*, 87(5), 1241-1262.doi: 10.1111/j.1540-6237.2006.00426.x.

Mahoney, K. S., & MacSwan, J. (2005). Reexamining identification and reclassification of English language learners: A critical discussion of select state practices. *Bilingual Research Journal*, 29(1), 31-42. doi: 10.1177/0265532217718600

Mahoney, K., Haladyna, T., & MacSwan, J. E. F. F. (2009). The need for multiple measures in reclassification decisions: A validity study of the Stanford English Language Proficiency Test. In G. Wiley, J. S. Lee, & R. W. Rumberger (Eds). *The education of language minority immigrants in the United States*. Bristol: Multilingual Matters (pp. 240-262).

Malabonga, V., Kenyon, D. M., Carlo, M., August, D., & Louguit, M. (2008). Development of a cognate awareness measure for Spanish-speaking English language learners. *Language Testing*, 25(4), 495-519. doi: 10.1177/0265532208094274.

Marôco, J., Gonçalves, C., Lourenço, V., & Mendes, R. (2016). *PISA 2015 – Portugal. Volume I: literacia científica, literacia de leitura & literacia matemática*. Lisboa: IAVE, I.P.

Mollborn, S., Fomby, P., & Dennis, J. A. (2012). Extended household transitions, race/ethnicity, and early childhood cognitive outcomes. *Social science research*, 41(5), 1152-1165. doi: 10.1016/j.ssresearch.2012.04.002

Nassaji, H. (2003). L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *Tesol Quarterly*, 37(4), 645-670. doi: 10.2307/3588216.

Organisation for Economic Co-operation and Development. (OECD) (2016). *Programme for International Student Assessment. PISA*.

Oh, J. S., & Fuligni, A. J. (2010). The role of heritage language development in the ethnic identity and family relationships of adolescents from immigrant backgrounds. *Social Development*, 19(1), 202-220. doi: /10.1111/j.1467-9507.2008.00530.x

Ong, A. D., Phinney, J. S., & Dennis, J. (2006). Competence under challenge: Exploring the protective influence of parental support and ethnic identity in Latino college students. *Journal of adolescence*, 29(6), 961-979. doi: 10.1016/j.adolescence.2006.04.010.

Peña, E., Bedore, L. M., & Rappazzo, C. (2003). Comparison of Spanish, English, and bilingual children's performance across semantic tasks. *Language, speech, and hearing services in schools*, 34(1), 5-16. doi: 10.1044/0161-1461(2003/001).

Pereira, K. M., & Ornelas, I. J. (2011). The physical and psychological well-being of immigrant children. *The Future of Children*, 195-218. Disponível em: http://www.futureofchildren.org/futureofchildren/publications/journals/article/index.xml?journalid=74&articleid=546

Portes, A., & Hao, L. (2002). The price of uniformity: Language, family and personality adjustment in the immigrant second generation. *Ethnic and racial studies*, 25(6), 889-912. doi: 10.1080/0141987022000009368.

Primi, R. (2003). Inteligência: avanços nos modelos teóricos e nos instrumentos de medida. *Avaliação psicológica*, 2(1), 67-77. Disponível em: ttp://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712003000100008&lng=pt&tlng=pt.

Pumariega, A. J., Rothe, E., & Pumariega, J. B. (2005). Mental health of immigrants and refugees. *Community mental health journal*, 41(5), 581-597. doi: 10.1007/s10597-005-6363-1

Saxena, S., Ambler, G., Cole, T. J., & Majeed, A. (2004). Ethnic group differences in overweight and obese children and young people in England: cross sectional survey. *Archives of Disease in Childhood*, 89(1), 30-36. doi: 10.1038%2Fijo.2014.171.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, 24(3), 355-390. doi: 10.1177/0265532207077205.

Scarcella, R. C., & Zimmerman, C. B. (2005). Cognates, cognition, and writing: An investigation of the use of cognates by university second-language learners. In A. Tyler, M. Takada, Y. Kim, & D. Marinova (Eds). *Language in use: Cognitive and discourse perspectives on language and language learning* (pp. 123-136). Georgetown: Georgetown University Press.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching,* 47(4), 484-503. doi: 10.1017/S0261444812000018.

Shin, S. K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing,* 22(1), 31-57. doi: 10.1191/0265532205lt296oa.

Sohn, S., & Wang, X. C. (2006). Immigrant parents' involvement in American schools: Perspectives from Korean mothers. *Early Childhood Education Journal,* 34(2), 125-132. doi:10.1007/s10643-006-0070-6.

Song, M. Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing,* 25(4), 435-464. doi: 10.1177/0265532208094272.

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, 11(3), 233-249. doi: 10.1080/15434303.2013.869815.

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing,* 22(2), 211-234. doi: 10.1191/0265532205lt301oa.

Van Tubergen, F., & Kalmijn, M. (2005). Destination-language proficiency in cross-national perspective: A study of immigrant groups in nine western countries. *American Journal of Sociology,* 110(5), 1412-1457. doi: 10.1086/428931.

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, 65(2), 390-416. doi:10.1111/lang.12105

Vandergrift, L., & Goh, C. C. (2012). *Teaching and learning second language listening: Metacognition in action.* Routledge.

Wang, Y., & Wang, J. Q. (2002). A comparison of international references for the assessment of child and adolescent overweight and obesity in different populations. *European journal of clinical nutrition,* 56(10), 973. doi: 10.1038/sj.ejcn.1601415.

Woodcock, R. W. (1991). *Woodcock Language Proficiency Battery-revised: WLPB-R.* Riverside Pub.