



Psicologia: Reflexão e Crítica

ISSN: 0102-7972

prcrev@ufrgs.br

Universidade Federal do Rio Grande do Sul
Brasil

Andriola Bandeira, Wagner

Funcionamento Diferencial dos Itens (DIF): Estudo com Analogias para Medir o Raciocínio Verbal

Psicologia: Reflexão e Crítica, vol. 13, núm. 3, 2000, pp. 475-483

Universidade Federal do Rio Grande do Sul

Porto Alegre, Brasil

Disponível em: <http://www.redalyc.org/articulo.oa?id=18813315>

- Como citar este artigo
- Número completo
- Mais artigos
- Home da revista no Redalyc

redalyc.org

Sistema de Informação Científica

Rede de Revistas Científicas da América Latina, Caribe, Espanha e Portugal

Projeto acadêmico sem fins lucrativos desenvolvido no âmbito da iniciativa Acesso Aberto

sociodemográfica dos sujeitos. Neste âmbito, sempre que se identifiquem grupos de sujeitos para os quais haja suspeita de diferenças nas pontuações obtidas no teste, se deve aplicar algum procedimento para a análise de potenciais vieses (Martínez Arias, 1997).

Hambleton (1989a), ao finalizar sua exposição sobre as várias definições de viés, constata que todas elas padecem de um problema comum: não consideram a necessidade de controlar a própria capacidade dos sujeitos na variável latente medida pelo item ou teste. Reside aqui a principal diferença entre os conceitos de *viés* e de funcionamento diferencial dos itens (*DIF*). O procedimento DIF trata de controlar a magnitude da variável latente (geralmente expressa pela letra grega θ) no item avaliado, isto é, os grupos são comparados com respeito às suas pontuações no item ou teste, considerando-se que suas magnitudes na variável latente (capacidades, habilidades ou aptidões) têm idêntico valor (Hambleton, 1994).

Alguns autores, entre os quais Camilli e Shepard (1994), insistem que os índices estatísticos utilizados na análise do DIF por si mesmos não proporcionam prova de viés, preferindo denominá-los de *índices de discrepância do item* ou de *funcionamento diferencial do item*. Segundo eles, este último termo engloba os diferentes procedimentos estatísticos para a detecção de um possível funcionamento diferencial, todavia, insistem em que o DIF não é sinônimo de viés, embora alguns autores creiam que sim. Desse modo, os termos *funcionamento diferencial do item* (DIF) e *viés* não deveriam empregar-se como sinônimos.

É que os métodos estatísticos de DIF se utilizarão para identificar itens que exibem um funcionamento diferencial nos distintos grupos e, posteriormente, depois de uma análise lógica ou experimental no contexto da validade de construto dos itens, se determinará quais têm vieses para que, assim, possam ser eliminados do teste ou banco de itens. Em outras palavras, os métodos de

Definição do Termo “Funcionamento Diferencial do Item” (DIF)

Atualmente, quase não se utiliza o termo *funcionamento diferencial do item* que foi preterido pelo de *funcionamento diferencial do item*. Uma das razões que explicam tal troca são as técnicas desenvolvidas para a detecção de DIF, proliferado muito nos últimos anos, nos Estados Unidos de América.

Outra razão deve-se à potência e flexibilidade para o estudo do DIF, já que foram desenvolvidos o objetivo de detectar se um item funciona de maneira diferente para grupos de distintas características sociodemográficas, cujos sujeitos compõem a mesma magnitude na variável latente (Penfield, 1997; Jiang & Stout, 1998; Kim & Oshima, Raju & Flowers, 1997; Scheunemann, 1997; Wainer & Lukhele, 1997; Williams, 1997). Além disso, as técnicas para a detecção do DIF, por mais além, quer dizer, não possibilitam a obtenção de informação acerca de sua natureza ou causa (Cohen & Baker, 1993; Muñoz, 1997).

Cohen, Kim e Baker (1993) distinguem o DIF quanto aos objetivos pretendidos: detectar ou verificar o impacto do DIF. No primeiro caso, as investigações que utilizam algum método para a identificação do DIF. Neste caso, os estudos buscam somente a detecção do DIF, quer dizer, se há uma possível diferença entre as CCI's dos itens para os grupos comparados. No segundo caso, as investigações para a busca e identificação do DIF. Neste contexto, o objetivo do investigador é determinar as distintas técnicas para a detecção do DIF, as quais são as causas (psicológicas, educacionais, sociais, atitudinais, etc.) que ocasionam o funcionamento diferencial de um item (Downing & Hambleton, 1996).

No âmbito da TRI um item não tem uma única curva característica do item (CCI) é idêntica para todos os grupos.

- T_{jGR} é a pontuação verdadeira do sujeito j que pertence ao grupo de referência e que tem uma certa magnitude na variável latente θ ;
- T_{jGF} é a pontuação verdadeira do sujeito j que pertence ao grupo focal e que tem uma certa magnitude na variável latente θ .

De acordo com Mazor, Hambleton e Clauser (1998), o uso do número de respostas corretas para a determinação do DIF só é aceitável no caso do teste ser unidimensional e, ademais, quando as respostas são dicotômicas. Como modo de visualizar o DIF em um hipotético item se apresenta a Figura 1.

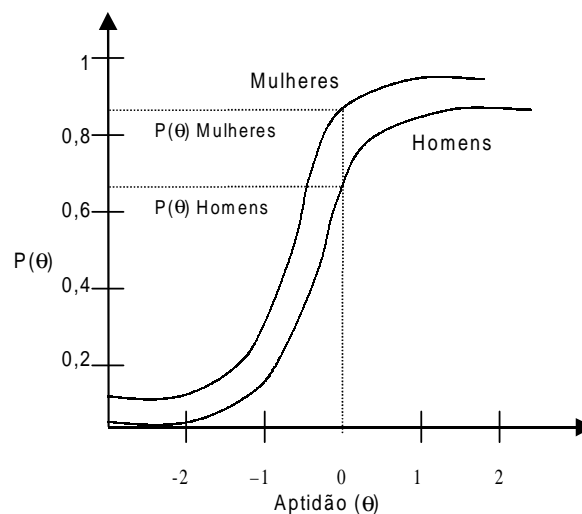


Figura 1. Representação gráfica das CCI's de um item com DIF

Podemos observar que para uma mesma magnitude de θ o valor de $P(\theta)$ é sempre superior para as mulheres, ou seja, em níveis iguais de competência na variável medida θ não existe a mesma probabilidade de superar o item. Neste caso, o item tem viés contra os homens (GR) pois

Ainda utilizando o item da Figura 1, vamos analisar o que ocorre com um item sem DIF, a Figura 2.

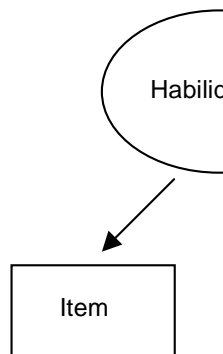


Figura 2. Relação entre habilidade e item sem DIF

O círculo indica a habilidade latente, que tem uma relação causal com o item. Em um teste com itens sem DIF, as habilidades latentes estão associados. Em um teste com itens com DIF, as habilidades latentes não estão associados. No modo de exemplificar, poderíamos dizer que uma pessoa com elevada habilidade na variável latente tem uma relação causal com o item. O grupo com maior capacidade de resposta na variável latente têm mais respostas corretas. O rendimento no item depende da habilidade latente, ou seja, se trata de um item sem DIF.

Agora, observemos o que ocorre com um item com DIF, representado na Figura 3.



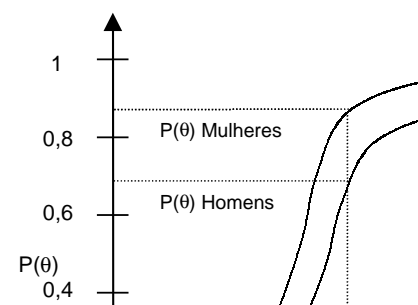
pode favorecer o rendimento de um grupo sobre outro devido a características particulares como sexo, raça, *background* educativo, origem social, etc. Deve ser enfatizado que, neste caso, se supõe que a magnitude da variável latente está sendo controlada, ou seja, os sujeitos são comparados com respeito a seu rendimento tendo em conta que possuem a mesma magnitude no construto. Este segundo exemplo caracteriza o caso no qual o rendimento no item não depende somente da magnitude da variável latente que os indivíduos tenham, senão de características do grupo, isto é, se trata de um item com DIF. Em nosso exemplo, a característica do grupo que influencia o rendimento diferenciado no item é o fato do sujeito ser homem ou mulher. Em síntese, se trata de uma característica de natureza demográfica que influencia o rendimento dos sujeitos com a mesma capacidade (Andriola, no prelo a).

É necessário reconhecer que o DIF ocasiona sérias implicações ao processo de avaliação, já que pode privilegiar um determinado grupo em detrimento de outro (Douglas, Roussos & Stout, 1996), conforme observamos no exemplo do rendimento dos homens e das mulheres. Muñiz (1997) adverte que o problema pode ter repercussões sociais mais graves se é a cultura dominante que elabora os itens para avaliar os sujeitos oriundos de culturas minoritárias. Por exemplo, suponhamos que são construídos itens para avaliar a capacidade de raciocínio verbal em alunos de escolas públicas e privadas. Ocorre que os alunos destes tipos de escolas são, geralmente, oriundos de classes sociais muito distintas, com diferentes *backgrounds* culturais, sociais, econômicos, etc. Todos estes aspectos podem implicar em que um tipo de aluno tenha o vocabulário mais rico que o outro. Dado que o raciocínio verbal é medido através de itens que empregam palavras, muito provavelmente, aquele tipo de aluno que conheça melhor o vocabulário utilizado nos itens terá uma clara vantagem na resolução destes mesmos itens.

Como nos fala Muñiz (1997), interminável e pode dizer-se que não é possível encontrar itens isentos completamente de vieses. No entanto, trata-se de detectar a quantidade de vieses que há em um determinado teste ou item. Finalmente, deve ser enfatizado que neste contexto, a importância é que objetivam verificar a existência de DIF, o que é justificada. Cabe ao avaliador verificar se há itens com DIF para que possa ser evitada a utilização de itens explicativos e, assim, evitar sua utilização em desvantagem (Hambleton, 1989b; L.

Tipos de DIF no Âmbito da TRI

Enfatizamos que no contexto da TRI a detecção do DIF consiste na comparação dos itens investigados, considerando os grupos de referência e focal. Os distintos métodos para a detecção de DIF foram desenvolvidos baseados em dois tipos de DIF. O primeiro tipo é denominado *DIF uniforme* e se observa quando as CCI's do item são paralelas com respeito aos grupos de referência e focal. O segundo tipo é denominado *DIF não uniforme* e se observa quando as CCI's não se cruzam. Em outras palavras, há uma vantagem relativa para um dos grupos de referência cujo valor é constante ao longo de toda a escala atitudinal. Este caso ocorre quando a discriminação (discriminação) tem o mesmo valor para os dois grupos, é, quando as CCI's são paralelas. Este caso é representado na Figura 4.



homens (grupo de referência), indicando que o item é mais fácil para o grupo focal. Tal diferença no parâmetro b (dificuldade) supõe que o item possui DIF. O segundo tipo de DIF é denominado *DIF não uniforme* ou *inconsistente*, e se observa quando as CCI's do item estudado, com respeito aos grupos de referência e focal, são diferentes e, ademais, se cruzam em algum ponto do intervalo atitudinal. Em outras palavras, quando há uma vantagem relativa para um desses grupos investigados, cujo valor é variável ao largo de todo o intervalo atitudinal. Este caso ocorre quando os parâmetros a , b ou c têm valores distintos nas duas CCI's, isto é, quando as CCI's não são paralelas. Este tipo de DIF está representado na Figura 5.

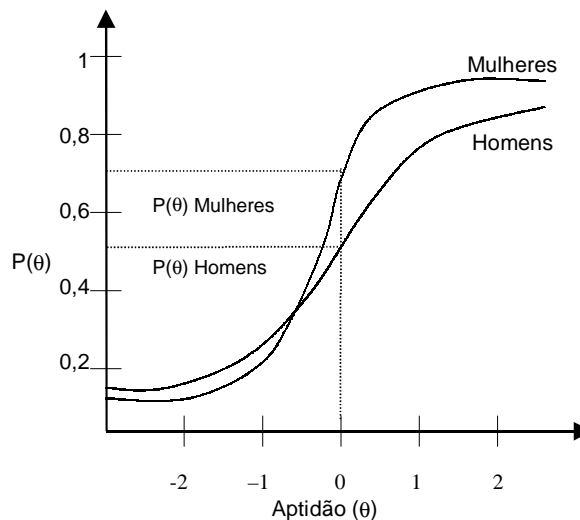


Figura 5. Representação gráfica de um item com DIF não-uniforme

A Figura 5 ilustra o caso de diferenças nos parâmetros a , b e c para os dois grupos investigados. É necessário dizer que neste segundo tipo de DIF é inapropriado um exame global dos dados, dado que tal procedimento poderia ocultar sua presença. É que a natureza variável

Swaminathan & Rogers, 1991) subjacente à detecção do DIF

- Estimar os parâmetros dos grupos de comparação;
- Colocá-los em uma mesma escala;
- Representá-los através das CCI's;
- Comparar os valores das CCI's dos grupos escolhidos.

Já dissemos que, segundo a definição de DIF, um item terá DIF se sua CCI não é a mesma para os dois grupos comparados, cujos sujeitos têm a mesma aptidão na variável latente θ (Kim, 1995). Esta definição tem como consequência desta definição que a CCI não é adequada para avaliar o DIF, pois ela não considera ambos os grupos e sua posteriori. Portanto, a CCI que difiram significativamente para um determinado item tem DIF (Mazor, Hambrecht, 1993).

Não obstante, o problema da detecção de DIF em uma investigação psicológica e educacional é complexo, com precisão a discrepância entre os resultados de distintos subgrupos. Assim, há vários procedimentos para o estudo de DIF (Mazor, Hambrecht, 1993; Holland e Wainer (1993), entre outros).

- O método das áreas;
- O método das probabilidades;
- O método de comparação das CCI's métricas dos itens;
- O método da regressão logística;
- O método do Qui-quadrado;
- O método de Mantel-Haenszel.

As investigações para a detecção de DIF estão baseadas em uma mesma escala. A detecção de DIF é um fator que pode influenciar a interpretação, realizada a partir de uma mesma escala, de um sujeito num item ou teste. É que a interpretação da pontuação, seja no âmbito psicológico ou educacional, toda a credibilidade e representatividade dos resultados dependem da validade da escala utilizada.

Método

Participantes

Foi composta por 730 estudantes do Ensino Médio, cuja idade média foi 17,7 anos ($dp = 3,12$ anos), sendo a maioria originária de escolas públicas (58,5%) e pertencendo ao sexo feminino (53,2%).

Instrumento

Foram utilizadas 30 analogias verbais componentes de um banco de itens já calibrados anteriormente para o uso com a população de estudantes do segundo grau, através do modelo logístico de dois parâmetros (Andriola, 1998).

Procedimento

Depois dos primeiros contatos com os dirigentes das escolas, para organizar os horários e as turmas que seriam utilizadas no estudo, os 30 itens foram aplicados de modo coletivo na amostra de estudantes. Não houve limitação de tempo para a resolução dos itens.

Resultados

Para a análise estatística dos dados utilizou-se o programa BILOG-MG, desenvolvido por Zimowski, Muraki, Mislevy e Bock (1996). Este programa utiliza o método de comparação dos parâmetros métricos dos itens para verificar a existência de DIF, mais especificamente a dificuldade do item (parâmetro b). Todavia, os itens utilizados no estudo foram calibrados através do modelo TRI de dois parâmetros logísticos, ou seja, tiveram determinados sua discriminação (parâmetro a) e dificuldade (parâmetro b). Nesse caso, o programa assume como iguais os valores do parâmetro a para os grupos de comparação (homens e mulheres).

Como afirma Martínez Arias (1997), o uso deste procedimento pretende testar as seguintes hipóteses:

Onde:

- $s\Delta_b$ é o erro-padrão da diferença entre os parâmetros b nos grupos focal e de referência;
- s^2_{bR} é a variância do parâmetro b no grupo de referência;
- s^2_{bF} é a variância do parâmetro b no grupo focal.

Neste âmbito, a prova de contraste é baseada na diferença dos parâmetros b pelo teste de Z, isto é:

$$Z = \frac{\Delta b}{s_{\Delta b}}$$

Onde:

- Z é o resultado da prova de contraste;
- Δb é a diferença entre os parâmetros b dos grupos focal e de referência;
- $s\Delta_b$ é o erro padrão da diferença entre os parâmetros b dos grupos focal e de referência.

Como Z tem uma distribuição aproximadamente normal, podem ser usadas as tabelas de distribuição normal para comprovar a significância do valor obtido na comparação desse valor com o valor crítico.

No nosso caso, o procedimento inicial para a análise do DIF consistiu na determinação dos parâmetros métricos dos itens para a amostra total (homens e mulheres) e o valor obtido para o teste da bondade de ajuste do modelo (método de máxima verossimilhança) foi 25.633,96. Em seguida, os parâmetros métricos foram determinados para os grupos de comparação (homens e mulheres). Deve ser dito que tais parâmetros devem estar em uma mesma escala. O teste da bondade de ajuste do modelo (método de máxima verossimilhança) para os dados, considerando o grupo de referência, resultou no valor 25.441,77.

A diferença entre os dois valores resultantes da bondade de ajuste do modelo tem uma distribuição *qui-quadrado* e, se é significativa, indica a existência de DIF no conjunto de itens. A diferença entre o valor inicial (25.633,96) e o valor resultante foi 192,19. Dissemos que esse valor é significativo se for maior do que o valor crítico, que é 192,19.

Tabela 1. Valores do Parâmetro b (dificuldade) dos 30 itens

Itens	Dificuldade dos itens (Parâmetro b)		Diferença entre os grupos (GF-GR)	Erro
	Homens (GR)	Mulheres (GF)		
1	3,469	3,303	- 0,166	
2	1,132	0,988	-0,145	
3	0,004	-1,401	-1,405	
4	-2,113	-2,327	-0,214	
5	-1,818	-1,460	0,358	
6	-1,264	-1,373	-0,109	
7	-1,302	-1,625	-0,323	
8	-0,844	-1,064	-0,220	
9	-1,209	-1,023	0,186	
10	-0,600	-0,145	0,455	
11	-1,066	-0,913	0,153	
12	-0,877	-1,625	-0,748	
13	-0,210	0,035	0,245	
14	-0,349	-0,021	0,328	
15	-1,101	-0,550	0,551	
16	-1,264	-1,105	0,159	
17	-0,827	-0,483	0,344	
18	-0,600	-0,927	-0,327	
19	-0,287	0,428	0,716	
20	-0,537	-0,671	-0,134	
21	-1,014	-1,091	-0,077	
22	-1,283	-1,358	-0,075	
23	-0,088	-0,430	-0,342	
24	-0,272	-0,281	-0,010	
25	-0,134	0,134	0,268	
26	-0,011	-0,159	-0,147	
27	0,019	0,368	0,349	
28	-0,241	-0,254	-0,013	
29	0,126	0,309	0,183	
30	-0,011	0,149	0,160	

Podemos observar inicialmente, que 16 itens (53,3%) — os parâmetros b dos grupos

Considerações Finais

O estudo relatado não tinha o objetivo de identificar as causas do DIF, mas verificar sua existência entre 30 analogias componentes de um banco de itens destinados à avaliação do raciocínio verbal em estudantes do ensino médio. Como afirmam Cohen, Kim e Baker (1993), se trata de uma investigação para detectar o DIF, isto é, determinar a possível diferença entre as CCI's dos itens, de acordo com os grupos comparados, adotando, para tanto, o critério de comparação do parâmetro b (dificuldade) dos itens.

Assim mesmo, como modo de verificar a plausibilidade de se tratar de itens multidimensionais que, supostamente, é uma causa de DIF (Andriola, no prelo b), observamos as cargas fatoriais destes cinco itens no fator único extraído através de análise fatorial (método de máxima verossimilitude). Ditas cargas se situaram entre 0,328 e 0,582, ou seja, não podemos afirmar que tais itens sejam multidimensionais, dado o elevado valor de suas saturações no fator extraído. Não obstante, haveria que tentar identificar outras possíveis causas do DIF destes cinco itens, adotando outros procedimentos, tais como, a análise de conteúdo de ditos itens por expertos na área.

É necessário afirmar que a principal contribuição desta investigação para a área da avaliação psicológica e educativa foi identificar os itens com DIF, que são componentes de um banco já organizado e pronto para ser utilizado na avaliação psicológica de estudantes do ensino médio. Os resultados possibilitarão, desse modo, que estes cinco itens não sejam utilizados nos processos de avaliação do raciocínio verbal em dita população de estudantes. Finalmente, a modo de conclusão, queremos apresentar algumas palavras de Pasquali (2000), destacando a importância dos estudos sobre o DIF:

“[...] essas informações sobre cada item de um teste [...] lhe permitem 1) tomar decisões sobre a qualidade do mesmo e 2) colocá-lo num banco de itens sem que ele perca a sua identidade, porque você tem uma série de indicadores

- Camilli, G. & Penfield, D. A. (1997). Variance estimation test functioning based on Mantel-Haenszel. *Educational Measurement*, 34, 123-139.
- Clauser, B. E., Nungester, R. J. & Swaminathan, H. (1998). Matching for DIF analysis by conditioning on educational background variable. *Journal of Educational Measurement*, 33, 453-464.
- Cohen, A. S., Kim, S. & Baker, F. B. (1993). Detecting differential functioning in the graded response model. *Educational Measurement*, 17, 335-350.
- Douglas, J. A., Roussos, L. A. & Stout, W. (1996). Item testing: identifying suspect bundles and assessing item functioning. *Journal of Educational Measurement*, 33, 123-125.
- Downing, S. M. & Haladyna, T. M. (1997). Test item evidence from quality assurance procedures. *Education*, 10, 61-82.
- Hambleton, R. K. (1989a). Introduction. *International Research*, 13, 123-125.
- Hambleton, R. K. (1989b). Principles and selected response theory. Em R. L. Linn (Org.), *Educational Measurement* (147-200). New York: MacMillan.
- Hambleton, R. K. (1994). Item response theory: a framework for measurement advances. *Psychological Measurement*, 1, 1-27.
- Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Item response theory*. North Carolina: Sage.
- Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum.
- Jiang, H. & Stout, W. (1998). Improved type I error rate estimation bias for DIF detection using SIBTEST. *Journal of Educational Measurement*, 23, 291-322.
- Kim, S. & Cohen, A. S. (1998). Detection of differential functioning under the graded response model with the Rasch model. *Applied Psychological Measurement*, 22, 345-355.
- Martínez Arias, R. (1997). *Psicométrica: Teoría de los tests*. Madrid: Síntesis.
- Mazor, K. H., Hambleton, R. K. & Clauser, B. E. (1991). DIF analyses: The effects of matching on unobserved scores. *Applied Psychological Measurement*, 22, 357-370.
- Mellenbergh, G. J. (1989). Item bias and item response theory. *Journal of Educational Research*, 13, 127-143.
- Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, 33, 379-416.
- Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Pirâmide.
- Muñiz, J. & Hambleton, R. K. (1992). Medio siglo de investigación a los ítems. *Anuario de Psicología*, 52, 41-66.
- Oshima, T. C., Raju, N. S. & Flowers, C. P. (1997). Demonstration of multidimensional IRT-based differential functioning of items and tests. *Educational Measurement*, 34, 253-272.

- Williams, V. S. L. (1997). The “unbiased” anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education*, 10, 253-267.
- Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (1996). *BILOG-MG. Multiple Group IRT Analysis and Test Maintenance for Binary Items*. Chicago: Scientific Software International (SSI).

Sobre o autor

Wagner Bandeira Andriola é Professor do Mestrado em Avaliação Educacional da Universidade Federal do Ceará (UFC); Psicólogo pela Universidade Federal da Paraíba (UFPB); Especialista em Psicometria pela Universidade de Brasília (UnB); Mestre em Psicologia Social e do Trabalho pela Universidade de Brasília (UnB); Doutorando do Programa *Investigación, Diagnóstico y Evaluación para la Calidad Educativa* da *Universidad Complutense de Madrid (UCM)*; Bolsista da Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Atua nas áreas de Instrumentação e Medida Psicológica e Educativa utilizando as Teorias Clássica dos Testes (TCT) e de Resposta ao Item (TRI), com especial interesse pelo estudo do funcionamento diferencial dos itens (DIF) em testes psicológicos e educativos.



Núcleo de Estudos e Capacitação em Desenvolvimento Humano

Objetivo Geral: Implementar a formação

Desenvolvimento Humano de profissionais, técnicos, agentes de comunidade e estudantes das áreas de Educação e Saúde, através de discussões teórico-temáticas, dinâmicas pedagógicas e produção de material pedagógico.

NECADEH - CEP-RUA

Instituto de Psicologia

Rua Ramiro Barcelos, 2600, Sala 104