Resumo Histórico sobre Testes de Inteligência

Por Hindemburg Melão Jr.

Em nossos 6.000 anos de história, podemos encontrar registros de instrumentos e métodos engenhosos para medir com acurácia os tamanhos dos objetos, desde os átomos até os planetas, métodos para medir as temperaturas das estrelas, as massas dos elétrons, as velocidades da luz e do som, a idade dos fósseis, a pressão atmosférica, o comprimento de onda das cores, as chances de duas pessoas serem parentes e as chances de um bilhete de loteria ser premiado, temos registros de descrições de métodos para medir muitas coisas, mas não encontramos nenhum que seja apropriado para medir a inteligência.

Se voltássemos ao século V de nossa Era, poderíamos encontrar os chineses aplicando exames com a finalidade de dividir a população em três castas, de acordo com a capacidade intelectual, para distribuir as funções e as responsabilidades segundo as habilidades de cada um. Esses testes consistiam basicamente em avaliar a capacidade mnemônica, a capacidade para interpretar os textos clássicos e a habilidade para escrever poemas (mais informações em http://www.eskimo.com/~miyaguch/ – História das sociedades de elevado QI). Nas melhores acepções da palavra "inteligência", esses provavelmente foram os primeiros testes cognitivos de que se tem registro.

Pitágoras, no século V a.C., testava os aspirantes a se tornarem seus pupilos, mas eram testes de conhecimentos geométricos e dificilmente poderiam ser considerados testes de inteligência. Luiz Pasquali, reportando-se a Dubois, afirma que os testes já eram usados na China 3.000 anos antes de Cristo, enquanto Anne Anastasi, reportando-se a Bowman, diz que os testes chineses foram usados durante cerca de 2.000 anos (não especifica de quando a quando), e também comenta rapidamente que foram usados exames na Grécia Antiga e na Europa Medieval. Mas em todos esses casos, tratavam-se de testes de cultura, com baixo teor de "inteligência fluida", de modo que os testes chineses do século V d.C. provavelmente foram os primeiros testes de inteligência, no sentido mais estrito do léxico.

No Ocidente, os trabalhos pioneiros foram realizados pelo médico Esquirol, que em 1838 associou os diferentes níveis de retardo mental aos diferentes níveis de fluência verbal. A idéia de Esquirol é correta, mas seria importante fazer duas distinções relacionadas ao gênero:

1 – Para a grande maioria dos traços cognitivos, inclusive a linguagem, os níveis de habilidade entre as mulheres se distribuem com uma dispersão claramente menor do que a dispersão observada entre os homens. Em outras palavras, para a maioria das habilidades cognitivas e personalógicas, o desvio-padrão para homens é maior do que para mulheres. Esse fato é amplamente documentado em cerca de 30 milhões de pessoas examinadas pelo SAT ao longo de várias décadas, além de 12.000 pessoas examinadas pelo D.A.T., conforme citado na página 115 do livro "Testagem Psicológica", de Lee J. Cronbach.

2 – As mulheres apresentam, em média, fluência verbal superior aos homens. De modo geral, as mulheres se saem melhor do que os homens em praticamente todas as atividades que exigem rapidez para tarefas simples, enquanto os homens se sobressaem em atividades que exigem profundidade para lidar com problemas complexos.

Seja como for, Esquirol deu um passo importante no sentido de conceber meios de estimar níveis mentais, pois ele tinha razão sobre haver correlação entre deficiências na linguagem e atraso mental, porém trata-se de uma correlação fraca.

O pai da Psicologia Experimental foi Francis Galton, e o pai dos testes psicométricos foi Alfred Binet. Galton, a partir de 1884, projetou os primeiros testes destinados a medir a inteligência, no entanto seus resultados não chegaram a ser satisfatórios. Ele acreditava que a inteligência poderia ser o resultado de um conjunto de características simples, como diâmetro da cabeça, velocidade dos reflexos, acuidade visual etc. Ele tinha alguma razão nisso, mas as correlações não eram tão boas como ele esperava. A correlação entre QI e tamanho da cabeça, por exemplo, é cerca de 0,4, enquanto a correlação entre QI e velocidade dos reflexos para tarefas elementares é cerca de 0,3, portanto ambas muito baixas para que possam corroborar sua hipótese.

Os primeiros testes de inteligência propriamente ditos, que deram origem aos modernos testes de QI, foram desenvolvidos por Alfred Binet, no início do século XX, e começaram a ser utilizados em 1904. Em princípio, Binet pretendia criar um instrumento que possibilitasse o diagnóstico objetivo de deficiências mentais, além de medir a gravidade da deficiência. Para tanto, Binet formulou vários questionários, que foram aplicados em grupos de crianças de diversas faixas etárias, e com isso nasceu o conceito de “idade mental”. Segundo Anastasi, reportando-se a T. H. Wolf, Binet não gostava do termo “idade mental” e preferia usar a expressão “nível mental”. Isso demonstra a refinada percepção de Binet, porque objetivamente o termo “nível mental” teria evitado diversas confusões que ocorrem em virtude do uso inadequado da expressão “idade mental”. Pela comparação entre idade mental (medida pelo teste) e idade cronológica (ou biológica), era possível saber se uma criança tinha desenvolvimento atrasado ou acelerado, e era possível ainda saber o quão avançada ou atrasada ela estava em comparação às outras crianças de mesma idade. Binet e seu colega nesse trabalho, Theodor Simon, compreenderam claramente as limitações do método que usavam, e não ousaram ir além desse ponto, sem antes desenvolver uma metodologia mais apropriada e ter uma compreensão melhor do assunto. No entanto, em 1906, na universidade de Stanford, Lewis Madison Terman publicou uma versão aprimorada dos testes de Binet, que prontamente foi reconhecida como a melhor bateria de testes de inteligência da época, e, em 1916, por sugestão de William Stern (em 1912), foi introduzido o conceito de “**quociente de inteligência**”, representando a idade mental multiplicada por 100 e dividida pela idade cronológica. Foi assim que nasceu o termo “**QI**”.

Primeira grande mudança no conceito de QI e nos processos de normatização:

Embora a sugestão de Stern tenha sido bem aceita por seus contemporâneos, todos os estudos desenvolvidos desde a época de Binet já revelavam que o desenvolvimento mental em função da idade cronológica não era linear, portanto a hipótese de Stern, desde a época em que foi proposta, não representava satisfatoriamente os dados experimentais. Ao que tudo indica, Binet já sabia dessas possíveis distorções bem antes de começarem a disseminar o conceito “inadequado” de QI, como representação da divisão da idade mental pela cronológica. O próprio Binet não apreciava sequer o termo “Idade Mental”; ele preferia “Nível Mental”. Mas Binet faleceu em 1911 e não havia quem contestasse o sistema proposto por Stern, por isso os procedimentos inadequados continuaram a ser usados durante algumas décadas, produzindo resultados distorcidos.

De acordo com Stern, uma criança com 5 anos e idade mental de 10 anos teria QI 200, e outra criança com 7 anos e idade mental de 14 também teria QI 200. Para Stern, tanto a criança de 5 anos quanto a de 7 anos, quando se tornassem adultas, provavelmente teriam QI 200. Essa idéia também foi abraçada por Terman e todos os expoentes mundiais da Psicometria naquela época, no entanto, quando essas crianças se tornam adultas, aquela que tinha QI 200 aos 7 anos tende a se tornar um adulto com QI em torno de 160, enquanto aquela que tinha QI 200

aos 5 anos tende a se tornar um adulto com QI 145. Isso é mau para a teoria, porque a invariabilidade do QI é uma das premissas em que o modelo teórico se sustentava. Essa distorção foi extensivamente confirmada por um estudo desenvolvido pelo próprio Terman, ao longo de mais de três décadas, com um grupo de 1528 crianças com QIs acima de 130. Nesse estudo, ele encontrou crianças com QIs acima de 200 num nível de abundância 10.000 vezes acima do que seria esperado numa população adulta, e a incidência comparativa era tanto maior se tanto mais jovens fossem as crianças, demonstrando que havia algo errado com o método adotado para determinar o QI. Por isso o conceito de QI como representação da divisão da idade mental pela cronológica foi substituído pelo conceito de QI em função da raridade. Essa mudança foi proposta por David Wechsler, na década de 1930. Ele também sugeriu que o antigo conceito de QI passasse a ser chamado *ratio-IQ*, enquanto o novo conceito se chamaria *deviation-IQ*. Algumas referências on-line:

http://sweb.uky.edu/~jcscov0/ratioiq.htm
http://www.psychologicaltesting.com/iqtest.htm
http://www.psych.usyd.edu.au/difference5/scholars/wechsler.html
http://www.wilderdom.com/intelligence/IQCautionsInterpretingIQ.html

Na década de 1940, o uso de *deviation-IQ* em lugar de *ratio-IQ* já estava generalizado e as escalas de QI passaram a ser construídas com base em níveis de raridade. Para isso, utiliza-se a inversa de uma distribuição normal cumulativa, e o QI é determinado pelo escore padronizado z que representa a porcentagem de pessoas com desempenho mais baixo que o daquela cujo QI se pretende determinar. Assim, uma pessoa situada no percentil 50 tem QI 100, pois ela está acima de 50% da população e abaixo de outros 50%, logo ela tem exatamente o QI médio, que é 100, por definição. Uma pessoa situada no percentil 84% tem QI 115 (numa escala com σ=15), porque, numa distribuição normal, 84% dos escores devem ficar abaixo de $+1\sigma$, e um QI de 115 está no nível $+1\sigma$ ($+1\sigma$ = 1 desvio-padrão acima da média). Uma pessoa situada no percentil 97,7% tem QI 130, porque 98% dos escores devem ficar abaixo de $+2\sigma$, e um QI de 130 está no nível $+2\sigma$. Uma pessoa situada no percentil 99,87% tem QI 145, porque 99,87% dos escores devem ficar abaixo de $+3\sigma$, e um QI de 145 está no nível $+3\sigma$. Por esse método, o QI de um sujeito é determinado com base na quantidade de pessoas de um determinado grupo que alcançam escores maiores ou menores do que ele. Se o grupo for constituído por uma população não-seleta, atribui-se à média desse grupo o QI 100. Se uma criança de 6 anos obtém um escore no topo de 2% de um grupo não-seleto constituído exclusivamente por crianças de 6 anos, então ela está no percentil 98 entre as crianças de 6 anos e seu QI é calculado em 130. Quando ela tiver 10 anos, provavelmente continuará no topo 2% das crianças de 10 anos, e quando se tornar adulta (mais de 20 anos), provavelmente continuará no topo 2% de adultos. A vantagem desse sistema de normatização é que na maioria das vezes o QI não apresenta variações significativas com a idade, tal como acontecia pelo método antigo de dividir a idade mental pela cronológica. Dos 6 anos até a idade adulta, o QI permanece quase inalterado, que é o que pretendia Stern, mas só foi satisfatoriamente alcançado por Wechsler.

O uso de escores padronizados com média 100 e desvio-padrão 15 se disseminou pelo mundo todo. Alguns continuaram adotando desvio-padrão 16 para adultos e 24 para crianças, com o intuito de preservar a similaridade com os escores do método antigo, mas já não se calcula o QI pela relação entre idade mental e idade cronológica. Em vez disso, os escores são calculados pelo método de Wechsler, com base nos níveis de raridade, usando média 100 e desvio-padrão que pode variar entre 15 e 24, dependendo da escala. É importante esclarecer que o fato de usar o método de Wechsler não implica que se use necessariamente a escala de Wechsler. Nos testes usados pela Mensa, por exemplo, tanto o teste de Cattell quanto o de Raven usam desvio-padrão 24, enquanto os testes usados em quase todas as outras

sociedades de alto QI preferem usar desvio-padrão 16. O sistema usado é o de Wechsler, mas as escalas são de Terman-Binet e Cattell. Para fazer a correspondência, basta usar a fórmula:

$QI_{SB} = 100 + (QI_{CT} - 100)/1,5$

QI_{SB} = QI pela escala Stanford-Binet, QI_{CT} = QI pela escala Cattell.

Muitas fontes informam que testes infantis têm desvio-padrão 24, enquanto testes para adultos têm desvio-padrão 16. Tal informação é incorreta por vários motivos:

1 – Não existe uma separação dicotômica entre "criança" e "adulto", e não faz sentido dar um salto brusco de 16 para 24 quando se passa dos 15,99 anos aos 16 anos. A transição do desvio-padrão 16 para 24 é suave. Em idades mais tenras, o desvio-padrão é maior do que 24. Há casos conhecidos de crianças com 3 anos e QI 300 ou 400, mas elas se tornaram adultos com QI 160 a 170. Isso sugere que em idades muito precoces o desvio-padrão pode ser 50, 60 ou até 70, e vai gradualmente diminuindo com a idade, até se estabilizar em torno de 16 pontos, por volta de 20 anos de idade. No caso de Erika Widsten, por exemplo, filha de nosso amigo Petri Widsten (membro de Sigma VI), teve QI 260 no TIG-NV, aos 8 anos de idade. Isso corresponde a um QI de raridade em torno de 180 e um QI de potencial em torno de 220.

2 – O desvio-padrão não é uniforme. Dentro de um mesmo grupo etário, observa-se que a quantidade de crianças (ou adultos) com QI muito alto ou muito baixo é maior do que seria esperada pela teoria. Isso acontece porque a distribuição empírica verdadeira não é perfeitamente gaussiana e começa a degringolar rapidamente fora do intervalo -2σ a $+2\sigma$. John Scoville fez um estudo interessante sobre isso (http://sweb.uky.edu/~jcscov0/ratioiq.htm), mas como ele não estratificou as crianças por faixa etária, não é possível tirar nenhuma conclusão segura de seu trabalho. A título de curiosidade, podemos dizer que ele formulou um método para determinar a variação do desvio-padrão em função do escore z. Talvez os valores que ele encontrou não estejam suficientemente acurados para que possam ser usados sem restrições, mas a idéia que norteou seu trabalho é legítima e representa uma contribuição importante para o estudo das variações na raridade de níveis cognitivos em função do próprio nível cognitivo. Mas a metodologia deveria levar em conta também as variações na raridade de níveis cognitivos em função da idade. Sem considerar o comportamento conjunto dessas duas variáveis, não é possível fazer inferências seguras.

3 – Os valores 16 e 24 foram medidos empiricamente na época dos testes de Terman. Entre adultos homens, encontrou-se desvio-padrão 16. Entre crianças com várias idades misturadas, encontrou-se desvio-padrão 24. Porém se o teste for padronizado pelo método de Wechsler, como nos casos do WAIS e WISC, ambos têm, por definição, desvio-padrão 15, embora o WAIS seja para adultos e o WISC para crianças.

Uma pequena contribuição para o aprimoramento no conceito de QI.

O sistema proposto por Wechsler continua sendo usado no mundo inteiro, mas ainda apresenta vários problemas. Dois são básicos e fáceis de serem resolvidos, pois dizem respeito apenas à terminologia. Trataremos dos problemas mais básicos primeiro: os termos *ratio-IQ* e *deviation-IQ* propostos por Wechsler são inapropriados. O termo *ratio-IQ* é redundante, porque *"ratio"* e *"quotient"* são sinônimos. Um termo mais apropriado é "*age-IQ*" e deveria informar a idade em que o escore foi obtido. Por exemplo; age-$IQ_{(12,25)}$ indicaria o QI obtido pelo método de divisão da idade mental pela cronológica, em teste aplicado em uma

criança de 12,25 anos (12 anos e 3 meses). Informar a idade em que o teste foi aplicado é fundamental para a correta interpretação do escore, conforme vimos acima. No caso de *deviation-IQ*, o problema é que a expressão diz respeito à **quantidade de desvios-padrão acima da média**, não a **raridade correspondente à quantidade de desvios-padrão acima da média**. O significado só seria o mesmo se a distribuição empírica fosse perfeitamente gaussiana, mas no mundo real isso nunca acontece. A escala para QIs proposta por Wechsler representa raridade com que surgem pessoas de determinado QI, portanto o termo correto é "*rarity-IQ*".

Diferentemente do *age-IQ*, não é tão importante que o *rarity-IQ* informe a idade da criança na época que o teste foi aplicado, porque as variações no *rarity-IQ* em função da idade são menores e não tão sistemáticas. Portanto as duas mudanças necessárias na nomenclatura são estas:

1. O *"ratio-IQ"* passa a se chamar *"age-IQ"* e informar, entre parêntesis e em sub-escrito, a idade em que o teste foi aplicado (idade com duas decimais).

2. O *"deviation-IQ"* passa a se chamar *"rarity-IQ"*.

Essa mudança semiológica foi proposta pela primeira vez em meu texto relativo à norma do Sigma Test, em setembro de 2003, e, para minha satisfação, a proposta foi amplamente aceita e muito elogiada. Alguns comentários que me enviaram sobre isso estão disponíveis na página sobre o Sigma Test.

Não é nossa intenção fazer uma análise exaustiva dos testes de Wechsler, por isso não discutiremos as dezenas de aprimoramentos que deveriam ser feitos nesses instrumentos. Mas há uma característica especialmente relevante e que está presente em todos os testes, não apenas nos de Wechsler. Trata-se do uso de escalas de intervalo ou de ordem, em vez de escalas de proporção. O uso de escalas intervalares ou ordinais torna impossível obter qualquer informação sobre a relação de proporção de potencial entre dois QIs. Permite apenas medir proporções de raridade de potencial. Por exemplo: pessoas com QI 150 são 10 vezes mais raras do que pessoas com QI 137. Mas quantas vezes as pessoas com QI 150 são mais "inteligentes" do que as pessoas com QI 137? Isso não é possível determinar com base no método atual. À parte a discussão sobre o que o teste de QI efetivamente mede, o fato é que mede algum tipo de habilidade, mas pelo método atual não é possível estabelecer uma relação de proporção dessa habilidade. Uma escala capaz de determinar proporções de inteligência com base nos escores de QI tem sido o "Santo Graal" da Psicometria nas últimas décadas. A solução para esse problema foi aplicada pela primeira vez no Sigma Test, em 2003, e descrita no texto relativo à norma de setembro de 2003.

Alguns dos aspectos indesejáveis nos testes de Wechsler estão relacionados à nomenclatura, à arbitrariedade na atribuição de escores, à pressuposição de não existir diferença de gênero e à pressuposição implícita de que as cargas fatoriais de todos os itens é a mesma em cada subteste, além de gerar uma escala com escores até 55, sendo que a padronização exclui de antemão pessoas portadoras de deficiências acentuadas (amostra clínica), portanto a norma não pode gerar escores corretos abaixo de 70. Muitas dessas falhas estão presentes não apenas nos testes de Wechsler, mas em todos os testes psicométricos. Mais adiante (em artigos futuros) discutiremos esses problemas e apresentaremos algumas soluções.

Segunda grande mudança no conceito de QI e nos processos de normatização.

Pouco depois dos trabalhos iniciais de Wechsler, surgiram os primeiros rudimentos de Teoria de Resposta ao Item, começando com Richardson, em 1936, recebendo importantes contribuições de Lawley, Lord e Birnbaum, e se consolidando com Georg Rasch, que em seu livro *"Probabilistic Models For Some Intelligence And Attainment Test"*, publicado em 1960, praticamente instaurou uma aliança entre a Matemática e a Psicologia, dando à Psicometria um aspecto mais robusto e mais científico. Nos anos seguintes, os avanços foram muito rápidos e significativos, em grande parte graças à velocidade de processamento proporcionada pelo advento do computador, alcançando seu apogeu com a popularização dos computadores pessoais, a partir da década de 1980.

Enquanto a "Teoria Clássica de Testes" (TCT) dá um tratamento conjunto a todos os itens que constituem cada prova, a Teoria de Resposta ao Item (TRI) faz uma análise individual de cada questão em comparação ao resto da prova. Isso possibilita obter um volume muito maior de informações relevantes e elimina diversas arbitrariedades da TCT. Por exemplo: em TCT usava-se *Split-Half* para calcular a homogeneidade de um teste. O método consiste em dividir o teste aleatoriamente em duas metades (questões pares e ímpares, por exemplo) e depois, calcular o coeficiente de correlação linear de Pearson entre essas metades e fazer o ajuste de Spearman-Brown. Dependendo das metades escolhidas, o resultado encontrado pode variar muito. As variações são pequenas, mas seria melhor obter um valor que não variasse e que representasse a média das correlações de todas as metades possíveis. É justamente isso que faz a TRI. Outras vantagens da TRI em comparação à TCT estão relacionadas à determinação da dificuldade de um item para diferentes níveis de habilidade e à estimação dos níveis de habilidade. Em TRI o conceito de escore padronizado z é inicialmente mantido, para determinação preliminar das propriedades dos itens, mas depois pode-se recalcular os escores usando métodos mais apropriados, por aproximações sucessivas, eliminando algumas limitações que impossibilitavam obter resultados fidedignos fora do intervalo de -2σ e $+2\sigma$. Com o uso de TRI, a faixa de segurança nos escores se amplia consideravelmente, chegando quase tão longe quanto o teto de dificuldade do teste permitir, ou seja, pode ficar entre -5σ e $+2{,}5\sigma$. A TRI também possibilita determinar a incerteza nos escores em determinados níveis, torna os processos de ancoragem mais versáteis, eficientes e precisos, entre outras coisas.

Por todos esses motivos, a TRI vem desfrutando um prestígio cada vez maior e ganhando a preferência de psicólogos e educadores no processo de normatização e validação de testes. Não obstante todas essas importantes virtudes, a TRI ainda apresenta algumas limitações conceituais e operacionais. A determinação do poder de discriminação (parâmetro a), de dificuldade (parâmetro b) e do fator sorte (parâmetro c) em cada item, mediante o ajuste simultâneo desses três parâmetros, é tarefa que só pode ser realizada por softwares especializados ou por pessoas que conheçam linguagens de programação e matemática pesada. Mesmo com tais softwares e tais conhecimentos, os resultados que o modelo logístico de 3 parâmetros fornece são conflitantes com as próprias definições dos parâmetros, conforme veremos no artigo em que discutimos os parâmetros da TRI. Outro problema é que a estimação dos níveis de habilidade para pessoas com escore 100% ou 0% tende a $+\infty$ ou $-\infty$, respectivamente, o que não procede. Além disso, a TRI não aborda diretamente o enunciado dos itens, mas apenas dá tratamento estatístico aos resultados brutos. Na página 116 do livro "Psicometria", de Luiz Pasquali, considerado por alguns psicólogos a maior autoridade do Brasil em Psicometria, encontramos uma recomendação do autor sobre o item 24 do teste TNVRA (de autoria do próprio Pasquali), em que ele afirma que esse item deveria ser desconsiderado (excluído do teste) porque apresenta carga fatorial menor do que 0,30. No entanto, o critério da carga fatorial, se for considerado isoladamente, conduz a decisões

erradas, porque existe correlação entre a carga fatorial e o nível de dificuldade de cada item. No caso do TNVRA, a correlação entre ***b*** (dificuldade) e a carga fatorial é -0,63, portanto os itens mais difíceis se mostram com cargas fatoriais distorcidamente menores (se comparados aos mais fáceis), conforme podemos ver na tabela abaixo:

b	**carga**
-0.62	0.44
0.25	0.75
0.29	0.79
0.18	0.71
0.59	0.44
-0.48	0.79
0.62	0.80
-0.33	0.73
0.19	0.85
-0.12	0.80
-0.47	0.72
-0.08	0.82
-0.46	0.77
-0.06	0.85
-0.88	0.78
0.62	0.67
0.55	0.71
-0.04	0.82
-0.44	0.87
0.24	0.69
0.28	0.83
0.49	0.74
-0.17	0.89
1.46	0.29
1.94	0.43
0.60	0.65
0.50	0.66
-0.44	0.64
1.66	0.36
1.27	0.57
0.24	0.70

Essa distorção só não aconteceria se a distribuição dos níveis de dificuldade dos itens fosse como uma gaussiana muito aproximadamente simétrica, e somente nesse caso o critério da carga fatorial poderia ser usado isoladamente para decidir se é conveniente ou não excluir itens com baixas cargas fatoriais. No caso do TNVRA, os itens mais difíceis (maiores parâmetros ***b***) estão muito mais distantes da média aritmética das dificuldades do que os itens mais fáceis, por isso ocorre essa distorção. Como o TNVRA tem apenas 4 itens com ***b>1***, eu manteria o item 24 porque ele contribui para melhorar a precisão e a discriminação nos níveis mais altos de habilidade, e o fato de esse item apresentar baixa carga fatorial é comparativamente de menor importância nesse caso. Se fosse removido, a precisão nas proximidades do teto cairia cerca de 25%, enquanto a consistência interna provavelmente aumentaria em menos que 3%. Não disponho dos dados brutos para calcular exatamente esse aumento na consistência interna, por isso minha estimativa se baseia em que a média aritmética das cargas fatoriais dos itens do teste fica 1,86% maior se for removido o item 24, e a média geométrica das cargas fatoriais dos itens do teste fica 2,67% maior se for removido o item 24. Supondo que o efeito observado na média das cargas fatoriais ao excluir o item seja semelhante ao que ocorreria na consistência interna, podemos concluir que haveria cerca de 2% de aumento na consistência interna mediante a exclusão do item 24, isso sem levar em conta que provavelmente também haveria uma pequena redução na consistência interna em virtude do efeito Spearman-Brown, sendo difícil estimar corretamente a dimensão desse efeito para a exclusão de um item isolado e que tem baixa carga, mas podemos supor que no final da história haveria cerca de 1% a 2% de aumento na consistência interna, e uma perda de 25% de precisão no teto. Portanto haveria um grande prejuízo na precisão nas proximidades do teto em troca de um pequeno ganho na "consistência interna". Uma mudança com essas implicações seria vantajosa para o teste? Claro que não. E aqui surge outro assunto que "daria muito pano pra manga": qual é a vantagem "real" que se pode atribuir a um teste que tenha maior α de Cronbach do que outro? Ou maior *Split Half* do que outro? O fato de um teste ser mais homogêneo na natureza dos traços medidos pode, em alguns casos, representar mais uma desvantagem do que uma vantagem. Um teste em que todas as questões sejam quase iguais, deve ter altíssimo α de Cronbach, no entanto seria preferível um teste que apresentasse maior variedade no conteúdo, cobrindo praticamente todo o espectro do traço medido. Um teste com séries numéricas, por exemplo, em que todas as séries sejam do tipo:

1) 3, 4, 5, 6, ?
2) 102, 103, 104, ?
3) 46, 47, 48, 49, ?
4) 12, 13, 14, 15, ?

Em que a regra é sempre somar 1, o α de Cronbach desse teste seria altíssimo, devido à grande homogeneidade dos itens, praticamente todos com mesmo conteúdo e mesmo nível de dificuldade. Esse teste seria pouco representativo da habilidade que ele propõe medir, "apesar de" ter um altíssimo índice de homogeneidade, aliás, justamente por ter um alto α de Cronbach é que seria pouco representativo. Por outro lado, se as séries envolvessem fatoriais, polinômios, decimais de π, além de questões com as quatro operações básicas, potenciação e logaritmos, teríamos um teste que mede o traço em questão de forma mais abrangente e numa maior gama de níveis de habilidade, o α de Cronbach seria menor, mas o teste seria mais representativo do traço que se deseja medir. Portanto, a menos que a finalidade do teste seja medir o desempenho num intervalo muito estreito de níveis de habilidade e com reduzidíssima variedade de conteúdo, seria preferível ter um α de Cronbach mais baixo do que renunciar a outras propriedades psicométricas muito mais importantes. Essas são apenas algumas das razões pelas quais eu acredito que a questão 24 do TNVRA deveria ser mantida.

Conforme vimos, a TRI é uma ferramenta poderosa e com muitas virtudes, mas ainda precisa de aprimoramentos em pontos importantes. Esse é um assunto complexo e extenso, por isso a abordagem que daremos enfatizará apenas os pontos mais importantes. Algumas inovações propostas serão:

- Método para calcular variações no parâmetro $\boldsymbol{c}$ em função do nível de habilidade.
- Flutuações estatísticas no parâmetro $\boldsymbol{c}$ em questões de múltipla escolha.
- Determinação dos parâmetros $\boldsymbol{a}$, $\boldsymbol{b}$ e $\boldsymbol{c}$ sem ajuste simultâneo e com maior fidedignidade do que pelo método tradicional.
- Determinação das incertezas nos escores individuais, em vez de usar mesma incerteza para todos os indivíduos que tenham mesmo escore, com ganho na acurácia.
- Alguns aspectos conceituais, como interpretação de cargas fatoriais e de coeficiente de precisão, conteúdo dos enunciados, nome dos testes etc.
- Alguns aspectos técnicos, como balanceamento dos pesos das questões em função dos níveis de dificuldade e qualidade das respostas. Na norma de outubro de 2004 do Sigma Test, serão apresentados alguns comentários sobre isso.

Atualmente a TRI é a ferramenta mais usada para padronização de testes e é claramente melhor do que suas predecessoras, contudo ainda há muito a ser aprimorado para que seja possível cumprir a difícil missão de medir a capacidade cognitiva de maneira menos arbitrária, mais rigorosa, mais acurada e mais significativa.

Testes de performance mental e testes de produção intelectual:

Voltando à história dos testes, vimos que os chineses já os utilizavam 1500 anos atrás. Até onde sabemos, esses foram os primeiros esforços no sentido de medir a inteligência, mas é provável que os testes sejam muito mais antigos do que isso, talvez mais antigos do que a própria escrita. A essa altura, o leitor poderia perguntar: *"Mas o quê esses testes estavam medindo? A inteligência?"* E nossa resposta é categórica: tanto os antigos testes chineses do século V d.C. quanto os testes modernos medem o desempenho da pessoa na realização de tarefas específicas. Acredita-se que os escores atribuídos aos desempenhos nesses testes correlatam positivamente com uma grandeza inapreensível que chamamos "inteligência". Os testes não medem a inteligência propriamente dita, mas medem o desempenho da inteligência enquanto ela se manifesta na resolução de problemas. Se o teste for estruturado de modo a exigir que a inteligência tenha uma participação ampla e profunda no processo de discriminação, ao ponto de imprimir de forma balanceada as suas virtudes globais no escore final, então o teste será representativo da capacidade intelectual global. Caso contrário, será baixo o coeficiente de correlação entre os escores obtidos no teste e o verdadeiro nível intelectual das pessoas testadas.

Em outras palavras, não existem testes capazes de medir a inteligência, mas existem testes que podem medir a **produção intelectual** e existem testes que medem a **performance intelectual**. Os que medem a produção intelectual servem para diagnosticar talentos como os de Einstein e Da Vinci, ou seja, gênios criativos e profundos, e em níveis não tão altos, medem também a habilidade de presidir a grandes empresas, gerenciar grandes volumes de informações, coordenar projetos complexos e outros processos complexos, enquanto os testes que medem performance intelectual servem para reconhecer prodígios de velocidade, de memória, de precocidade, de conhecimento etc. Uma pessoa que consegue pontuar alto no Sigma Test é criativa e tem pensamento profundo. Uma pessoa que consegue pontuar alto no Cattell III, RAPM II ou WAIS III é uma pessoa rápida. Claro que algumas pessoas podem ser rápidas e também profundas, e quando isso acontece, que são casos raros, os testes de performance conseguem prognosticar corretamente a capacidade de produção. Em todos os outros casos, os prognósticos feitos com base em testes de performance são inverossímeis. Dos 1.528 casos estudados por Terman, de crianças com escores entre 130 até mais de 200, em testes de performance, nenhuma se mostrou genial em idade adulta, embora todas tenham apresentado bom desempenho profissional. Por outro lado, a maioria das crianças com alto desempenho em Olimpíadas da Matemática acabam se tornando adultos geniais, porque as Olimpíadas da Matemática medem mais produção intelectual do que performance intelectual. As provas dessas olimpíadas têm prazo, o que distorce a medida, no entanto o nível de dificuldade, o conteúdo e o tipo de pensamento necessário para resolver as questões é muito mais adequado para medir produção intelectual, por isso os prognósticos são muito mais bem sucedidos. Para fazer prognósticos tão bons quanto possível, é necessário utilizar testes que, pela forma e pelo conteúdo, atendam ao propósito de medir a capacidade de produção intelectual propriamente dita. O Sigma Test, o Sigma Test VI e o Sigma Test Light foram construídos com esse propósito e, até onde podemos avaliar, são os mais adequados para isso.

Os testes de inteligência modernos podem ser divididos em dois grandes grupos: os *home tests* e os *supervised tests*.

Home Tests:

Alguns exemplos de *home tests* são o Sigma, o LAIT e o Titan. Eles não têm limite de tempo, as questões não são explicitamente de múltipla escolha (embora algumas o sejam implicitamente) e os níveis de dificuldade são claramente mais elevados do que os *supervised*

tests, podendo chegar a medir corretamente QIs de potencial acima de 240 e QI de raridade até 190. Os *home tests* medem "capacidade de produção intelectual", porque o fato de não terem limite de tempo e as questões serem difíceis, a pessoa examinada precisa ter disciplina, organização, persistência, capacidade de permanecer longas horas concentrada no mesmo assunto, sem se dispersar, que são habilidades muito mais intimamente ligadas à capacidade de produção do que a simples rapidez para resolver problemas elementares. Aqui convém citar uma frase de Thomas Edison: *"o sucesso depende em 1% de inspiração e 99% de transpiração"*, podemos complementar essa frase dizendo que esses 1% de inspiração são importantíssimos e fazem toda a diferença, mas não são suficientes para alcançar o sucesso. Um gênio verdadeiro, como Einstein ou Newton, não é alguém que se levantou numa bela manhã e teve todas as idéias sobre como interpretar o universo. Em vez disso, eles trabalharam intensamente, pensando sobre os assuntos que investigavam praticamente o dia inteiro, a tal ponto que entravam num estado de completo envolvimento com a questão, e muitas vezes sonhavam com o tema estudado. O resultado natural era a emergência de idéias que ninguém havia tido antes, simplesmente porque ninguém havia se aprofundado tanto naqueles assuntos. Eles combinaram uma capacidade excepcional para compreender, para criar, para estabelecer todo gênero de analogia e associação, para analisar e criticar e muitos outras habilidades cognitivas, e combinaram com vários traços de personalidade, tais como uma interminável e inabalável motivação para o trabalho. Se tivessem apenas habilidades mentais, mas não o perfil de trabalhadores compulsivos, não teriam chegado tão longe. Portanto os *home tests* são instrumentos que medem muito melhor algumas das características mais importantes para a capacidade de produção intelectual, combinando traços cognitivos e personalógicos.

Supervised Tests:

Alguns exemplos de *supervised tests* são o WAIS, o Binet, o RAPM e o Cattell. Eles têm limites de tempo e esses limites podem ser para o teste todo ou para cada questão. Alguns desses testes têm questões de múltipla escolha, outros têm questões discursivas, outros possuem uma mistura das duas. Os níveis de dificuldade nunca ultrapassam o teto de QI 135 a 140, por isso só podem medir com segurança até o nível de QI 130 a 135. São testes que medem a "performance mental", e uma medida performática é menos representativa de uma característica que acreditamos ser relativamente estável, como é o caso da inteligência. Se hoje uma pessoa tem QI 118, não é aceitável que amanhã ela tenha QI 102. No entanto, é perfeitamente aceitável que num teste performático uma pessoa tenha escore 118 hoje e 102 amanhã, porque pode ser que hoje ela estava bem disposta, e amanhã com sono. Como os *home tests* não têm prazo para conclusão e podem ser resolvidos em várias sessões, a pessoa pode trabalhar nos problemas quando se sentir mais motivada, e se num teste feito em 2002 ela teve escore 118, é quase impossível que em 2003 ela tenha escore 102, porque fatores como cansaço, estresse, cefaléias ou qualquer outro que possa interferir no desempenho, é eliminado ou minimizado nos *home tests*. Um dos grandes disparates dos *supervised tests* e que leva os leigos em Psicometria a duvidar da validade de testes em geral é que pessoas como Richard Feynman, ganhador do Nobel de Física e um dos maiores expoentes intelectuais do século XX, obteve escore 123 num teste de QI, enquanto Adam Konantovich, Keith Ranieri e Marilyn vos Savant, que são pessoas muito inteligentes, mas não apresentam evidências de serem tão ou mais inteligentes do que Feynman e seus níveis de produção intelectual são muitíssimo mais baixos que os de Feynman, tiveram, respectivamente, escores 268, 242 e 228 em testes de QI. Em estimativas subjetivas, os quatro estariam situados aproximadamente no mesmo nível, oscilando num intervalo de 10 a 20 pontos, e eu arriscaria situar Feynman no topo dos quatro. Se Feynman pontua 123 num teste e essas pessoas pontuam acima de 220, o problema só pode estar no teste, não em Feynman, cujo potencial de criação é fartamente comprovado. Mais adiante, discutiremos a diferença entre desenvolvimento precoce e talento,

então ficará mais claro porque esses escores na infância são enganadores quando encarados como preditores de genialidade.

Comparação entre *Home Tests* e *Supervised Tests*:

Faremos, a seguir, uma comparação entre as vantagens dos *supervised tests* e as vantagens dos *home tests.* As principais vantagens dos *supervised tests* são estas:

1 – Um teste como o RAPM pode ser aplicado coletivamente, em milhares de pessoas, em apenas 40 minutos, e o resultado pode ser gerado eletronicamente, em questão de segundos. Essa qualidade está presente em vários *supervised tests* e os torna muito ágeis, fáceis de aplicar, fáceis de corrigir.

2 – A supervisão diminui os riscos de fraudes. Nos *home tests*, uma pessoa pode não respeitar as instruções e consultar material ou pessoal que não seja permitido. O Mega Test, por exemplo, faz uma série de restrições quanto ao uso de *search engines* e computadores, mas como não é possível ter controle sobre as pessoas examinadas, essas restrições acabam sendo encaradas por muita gente como piada e ou como defeito no teste, pois sem supervisão, as pessoas podem desrespeitar impunemente as regras. Esse defeito é minimizado no Sigma Test e é totalmente corrigido no Sigma Test Light, porque, no caso do Sigma Test, não há nenhuma restrição ao uso de qualquer material de pesquisa ou computadores; a única restrição é sobre consultas a terceiros, o que diminui consideravelmente o risco de fraude, porque além da pessoa faltosa precisar da colaboração de um cúmplice, seria necessário que esse cúmplice tivesse um alto nível de habilidade. Digamos que a probabilidade de a pessoa encontrar um cúmplice seja 5%, então ao multiplicar isso pela probabilidade desse cúmplice ter habilidade acima de QI 125 teríamos 0,05 x 0,05 = 0,0025, ou seja, 0,25% (isso supondo que a probabilidade de encontrar um cúmplice fosse igual para todos os níveis de habilidade). Logo, o risco de uma pessoa conseguir benefícios ilícitos no Sigma Test é muito menor do que o risco dessa pessoa ganhar 15 pontos de QI num teste de múltipla escolha, devido às flutuações estatísticas (acertos casuais). Ou seja, o risco de alguém obter escore no Raven inflacionado em 15 pontos devido à sorte é maior do que o risco de alguém ganhar 15 pontos ilícitos no Sigma Test. No caso do Sigma Test Light, o risco é eliminado porque o teste pode ser aplicado com supervisão, em várias sessões.

Agora vejamos algumas vantagens dos *home tests*:

1 – As possibilidades de erros acidentais são minimizadas pelo fato de não haver limite de tempo. Assim as pessoas mais meticulosas, que analisam profundamente um problema, não são prejudicadas por isso, como pode acontecer em testes de QI convencionais. Se alguém encontra a resposta certa, mas acha que foi muito fácil e continua analisando o problema à procura de uma solução mais exata, essa pessoa acaba sendo prejudicada num teste com limite de tempo (*supervised test*), e, o que é pior, ela será prejudicada por ter uma característica que deveria ser considerada uma virtude do ponto de vista cognitivo, pois revisar as respostas é uma qualidade, não um defeito. Portanto essa é uma vantagem muito importante dos *home tests* em comparação aos *supervised tests*, que permite revisar o teste várias vezes, incrementando o escore, sem riscos de esgotar o tempo.

2 – No caso específico e exclusivo do Sigma Test, as pontuações atribuídas a cada questão são proporcionais ao grau de dificuldade. Isso tem implicações de capital importância, porque se uma pessoa acerta as questões mais difíceis e erra as mais fáceis, pode-se concluir que seus erros nas mais fáceis foram por distração, enquanto os erros nas mais difíceis foram devidos à dificuldade intrínseca dessas questões. Logo, quem acerta as mais difíceis merece

uma pontuação maior. Isso refina a acurácia do teste. Além disso, as dificuldades não são determinadas arbitrariamente. São calculadas objetivamente e com base em bons critérios, conforme será descrito no capítulo relativo à escoragem.

3 – As questões mais difíceis são de fato extremamente difíceis e capazes de discriminar corretamente em todos os níveis que o teste propõe discriminar. Não se pode usar questões baseadas em regra de três composta, em álgebra elementar ou séries de figuras para mensurar QIs acima de 130, mas os testes de QI convencionais fazem justamente isso, e o que é ainda pior, usam tais questões para determinar QIs acima de 180, como nos casos do Binet IV e do Cattell III. A tabela abaixo mostra os tetos teóricos do Cattell III e do WAIS III, em comparação a outros parâmetros intelectuais.

	Binet QI	Wechsler QI	z	Percentil	Raridade (1/x)
Inteligência média	100	100	0	50,000%	2
Inteligência acima da média	116	115	1	84,134%	6
Superdotado, cut-off em Sigma e na Mensa	132	130	2	97,725%	44
Média dos engenheiros da NASA e do MIT	144	141	2,75	99,702%	336
Gênio, cut-off em ISPE	148	145	3	99,865%	741
Média dos ganhadores de Nobel em Ciência	155	156	3,75	99,991%	11.307
Teto teórico do WAIS	164	160	4	99,9968%	31.560
Cut-off em Sigma V e Pars Society	180	175	5	99,999971%	3.483.046
Teto teórico do Cattell III	187	181,56	5,44	99,9999973%	36.927.646
Cut-off teórico em Sigma VI e Giga Society	196	190	6	99,99999990%	1.009.976.678

Não é possível usar um teste com séries de figuras, como o Cattell III, para medir o desempenho num nível de habilidade muito acima da média dos ganhadores de prêmio Nobel em Ciência. Qualquer ganhador de prêmio Nobel pode estourar o teto desses testes com os olhos fechados e ainda por cima alcoolizados, desde que disponham de tempo suficiente. Mas não é aí que está a maior falha. O problema é que usar séries de figuras para medir a inteligência de um prêmio Nobel é o mesmo que tentar medir a inteligência de um escritor como Machado de Assis com base na velocidade com que ele resolve questões de soletrar palavras ou para resolver palavras cruzadas. A capacidade para resolver palavras cruzadas tem relação com a inteligência e pode medir corretamente a capacidade mental nos níveis de desempenho entre QI 70 e 130, mas um escritor como Machado de Assis teve habilidades extraordinárias, muito mais importantes e mais representativas de sua inteligência do que a velocidade para solucionar palavras cruzadas, habilidades que outras pessoas não possuem e que as palavras cruzadas não medem. Um teste que pretendesse medir sua capacidade deveria partir de características encontradas em Dostoiévsky, Kafka, Goethe, Shakespeare, Homero, Russell, Sartre, Voltaire e outros grandes expoentes da Literatura. Então esse teste deveria ser capaz de medir em que proporção essas habilidades estão presentes em cada um deles. Se não for dessa maneira, o resultado será um completo disparate e não faria nenhum sentido dizer que Dostoiévsky foi mais inteligente do Kafka porque soletrava palavras mais rapidamente ou que porque era melhor em palavras cruzadas. Para crianças de 8 anos, o uso de testes de soletrar pode correlatar bem com a capacidade da criança porque mede um tipo de habilidade que representa o teto de capacidade dela, mas não serve para nada se for usado em grandes escritores, porque o teto de habilidade deles envolve habilidades cognitivas de natureza muito diferente daquela medida pelos testes de soletrar. Portanto os testes convencionais não são bons nem ruins. Eles podem ser bons enquanto forem aplicados nas condições que atendam aos seus limites de aplicabilidade, e passam a ser ruins a partir do momento que esses limites são excedidos, porque eles começam a fornecer escores fictícios, cada vez mais altos, mas esses escores não significam aquilo que pretendem significar, ou seja, um QI 160 no WAIS não significa nada além de um QI de 130, mas um QI 160 no Mega Test ou no Sigma Test realmente significa um QI em torno de 160, porque o conjunto de

habilidades medidas e o nível de dificuldade são condizentes com o nível em que os testes propõem medir. Esses comentários são importantíssimos, porque existem centenas de testes cujo teto verdadeiro é 130 ou 135, mas que fornecem escores 150, 160 e até 187.

4 – No caso específico do Sigma Test, houve o máximo empenho para que as questões não exigissem nenhum conhecimento além do Ensino Médio, e a maioria dessas questões pode ser resolvida apenas com conhecimento de Ensino Fundamental. O Sigma Test não é tão isento de cultura quanto o Raven ou o Cattell, mas é seguramente menos influenciado pela cultura do que o Binet ou o WAIS. Isso é importante porque embora a cultura e o vocabulário apresentem boa correlação com a inteligência no intervalo entre 70 e 130, essa correlação se torna muito fraca para QIs acima de 130 e desaparece totalmente para QIs acima de 145. Então, para minimizar os ruídos e distorções, um bom teste que pretende discriminar corretamente em níveis acima do percentil 98 precisa exigir apenas informações que sejam seguramente dominadas por todas as pessoas examinadas. Portanto o Sigma Test produz escores mais confiáveis quando aplicado em pessoas com pelo menos ensino médio completo (School).

Vimos aqui uma comparação generalizada dos *home tests* e *supervised tests*. Agora vamos comparar alguns testes específicos, enumerando as variáveis medidas por cada um deles.

O que mede o Sigma Test?

É um teste baseado em problemas da vida real, que exigem criatividade, engenhosidade, planejamento, organização, pensamento analítico, pensamento lógico profundo e complexo, pensamento lógico simples e superficial, sensibilidade para perceber sutilezas. Esquematicamente podemos dizer que o Sigma Test mede as seguintes capacidades:

- Capacidade para compreender o pensamento de outras pessoas em questões profundas.
- Capacidade para compreender o pensamento de outras pessoas em questões superficiais.
- Capacidade para compreender a Natureza em questões profundas.
- Capacidade para compreender a Natureza em questões superficiais.
- Capacidade para compreender a Lógica em questões profundas.
- Capacidade para compreender a Lógica em questões superficiais.
- Capacidade para interpretar textos simples e complexos.
- Capacidade para perceber sutilezas evidentes em questões lógicas e situações da vida real.
- Capacidade para perceber sutilezas recônditas em questões lógicas e situações da vida real.
- Capacidade para usar rudimentos de Matemática elementar para resolver problemas simples do cotidiano em diferentes níveis de dificuldade.

O Sigma Test Light tem formato semelhante e pretende medir as mesmas habilidades do Sigma Test, além de algumas outras mais básicas, como capacidade para seguir instruções, aprendizado elementar etc.

O que mede o Sigma Test VI?

É uma versão mais difícil do Sigma Test. É baseado em problemas da vida real que exigem criatividade, engenhosidade, pensamento lógico profundo e complexo, sensibilidade para perceber sutilezas e capacidade para descobrir e inventar soluções engenhosas, elaboradas e eficientes para problemas de nível olímpico. Esquematicamente podemos dizer que o Sigma Test VI mede as seguintes capacidades:

- Capacidade para compreender o pensamento de outras entidades conscientes.
- Capacidade para compreender a Natureza e os limites do que é possível.
- Capacidade para compreender a Lógica em questões de nível olímpico.
- Capacidade para interpretar a mente e a natureza.
- Capacidade para formular soluções pertinentes que envolvam estruturas complexas.
- Capacidade de usar rudimentos de Matemática para resolver problemas de nível olímpico.

O que medem o Stanford-Binet e o WAIS-verbal?

São testes baseados em perguntas acadêmicas triviais, exigem atenção para evitar descuidos, pensamento lógico simples, superficial e rápido. Esquematicamente podemos dizer que medem as seguintes capacidades:

- Capacidade para ler e escrever.
- Capacidade para conhecer o significados de algumas palavras e reconhecer sinônimos.
- Capacidade para ter adquirido algumas informações variadas.
- Capacidade para ler e interpretar textos.
- Capacidade para aplicar fórmulas matemáticas simples na resolução de problemas elementares.
- Capacidade para perceber diferenças evidentes entre coisas semelhantes.
- Capacidade para perceber semelhanças evidentes entre coisas diferentes.
- O WAIS mede também a memória em curto prazo num nível muito básico.

O que medem o Cattell e o RAPM?

São testes baseados em séries de figuras, exigem atenção para evitar descuidos, pensamento lógico simples, superficial e rápido. Esquematicamente podemos dizer que medem as seguintes capacidades:

- Capacidade para perceber padrões de mudanças em desenhos simples.
- Capacidade para perceber tendências, com base nas mudanças nos desenhos simples.
- Capacidade para deduzir mudanças futuras, com base nas tendências observadas.
- Capacidade para estabelecer analogias simples entre figuras.

Conforme podemos perceber, o conjunto do que o Sigma Test mede é mais abrangente e mais profundo do que o conjunto do que é medido pelos testes tradicionais. Além disso, as perguntas são muito mais semelhantes a situações da vida real, por isso medem muito melhor a habilidade para lidar com problemas reais. Como o Sigma Test Light segue o modelo do Sigma Test, é esperado que tenha propriedades semelhantes.

Precocidade e talento:

É muito comum assistirmos, em programas de variedades, a crianças que memorizaram as bandeiras dos países, ou as capitais, ou qualquer outra informação, e essas crianças serem consideradas “gênios”. Mas todo psicólogo sabe que a genialidade não pode ser determinada com base apenas na memória, por isso nenhum profissional de Psicologia cometeria o erro de diagnosticar uma criança como gênio apenas por ela apresentar uma boa memória, e menos ainda por apresentar uma memória normal bem treinada. No entanto, há um erro sistemático e institucionalizado que é cometido quando se considera que uma criança precoce será um adulto genial, e outro erro sistemático e institucionalizado que é cometido quando, ao constatar que crianças de QI altíssimo não se tornam adultos geniais, presumir que nenhum teste poderia predizer a genialidade e que a limitação está no teste. O fato é que os testes tradicionais realmente não servem para predizer a genialidade e só acertam muito raramente no diagnóstico. Isso não significa que se fossem usados testes melhores, o problema continuaria insolúvel. O que os fatos nos revelam é que os gênios como Da Vinci, Pascal, Newton, Galileu, Einstein e outros, demonstravam sua genialidade desde a infância, mas não pela resolução veloz de questões primárias de aritmética ou séries de figuras. Eles demonstravam sua genialidade resolvendo problemas complexos de Lógica e Física, inventando instrumentos, desenvolvendo métodos matemáticos, concebendo teorias bem articuladas e formulando experimentos para testar a validade dessas teorias. Esses traços cognitivos eram apresentados desde os primeiros anos de vida, mas os testes tradicionais não medem isso, portanto é natural que esses testes sejam incapazes de fazer diagnósticos corretos sobre genialidade. Conforme já vimos, os testes de QI foram criados para predizer deficiências, e cumprem bem essa função. Usar os mesmos testes, sem as devidas alterações, com o propósito de identificar e medir talentos, é um erro grave e de sérias conseqüências. O Sigma Test é totalmente diferente e tem conteúdo que engloba desde questões básicas de pensamentos primitivos até questões complexas que exigem pensamentos requintados. Isso o torna muito apropriado para diagnosticar talentos.

Algumas definições que continuam vigentes até hoje usam o termo “gênio” em contextos inadequados. As nomenclaturas propostas por Terman e Wechsler, por exemplo, continuam servindo como referência em muitas clínicas e no livro “Psicometria”, de Luiz Pasquali, publicado em 2003, ele usa a classificação de Terman, mudando apenas detalhes na nomenclatura:

Terman's classification (σ=16)

IQ Range	Classification
140 and over	Genius or near genius
120-140	Very superior intelligence
110-120	Superior intelligence
90-110	Normal or average intelligence
80-90	Dullness
70-80	Borderline deficiency
Below 70	Definite feeble-mindedness

Wechsler's classification (σ=15)

Classification	IQ Limits	Percent Included
Very Superior	128 and over	2.2
Superior	120-127	6.7
Bright Normal	111-119	16.1
Average	91-110	50
Dull Normal	80-90	16.1
Borderline	66-79	6.7
Defective	65 and below	2.2

Fonte: http://members.shaw.ca/delajara/IQBasics.html

Conforme podemos observar, na escala de Wechsler os termos “gênio” e “talento” foram suprimidos, demonstrando que ele tinha uma melhor compreensão sobre os limites do que os testes podiam medir. Mas no caso de Terman, cujas definições ainda são usadas em muitos livros, cursos e clínicas, ora aparece o termo “gênio”, ora o termo “superdotado”, ora o termo “talentoso”. Não temos objeção a fazer sobre a nomenclatura de Wechsler, mas há erros na

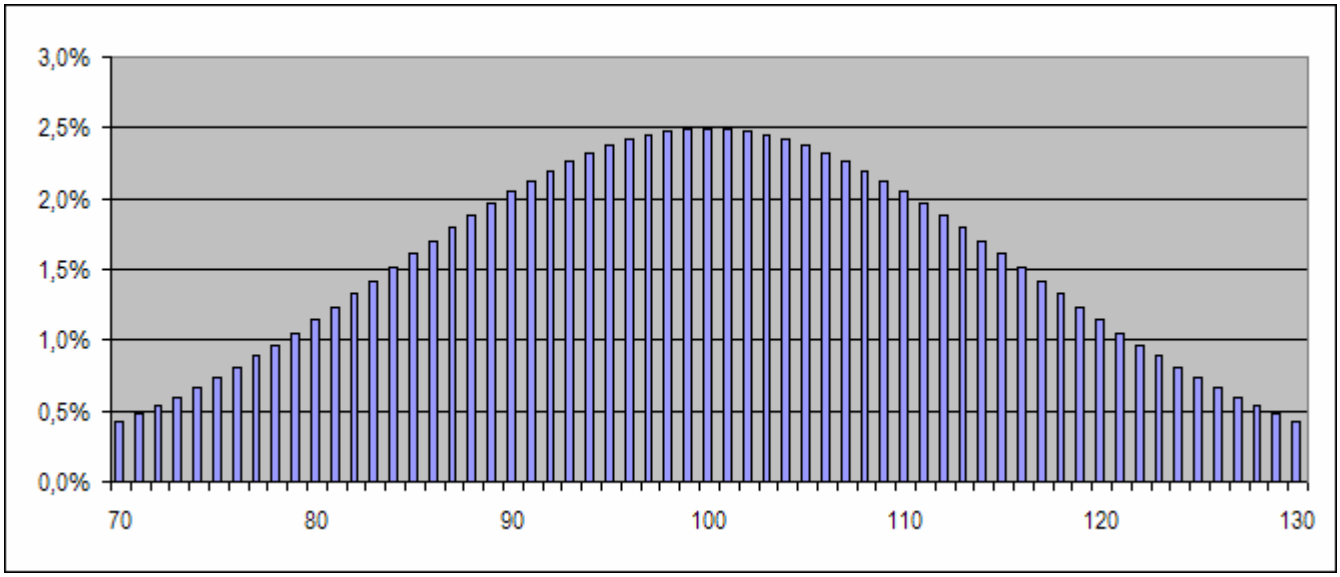
nomenclatura de Terman, porque uma criança de 5 anos que tem desempenho igual ao da média das crianças normais de 10 anos é uma criança precoce, é um prodígio infantil, mas o fato de ela ter apresentado um desenvolvimento acelerado não é necessariamente um indício de genialidade, porque as habilidades exigidas para atuar tão bem quanto a média das crianças de 10 anos não nos oferece nenhum dado concreto que nos permita prognosticar que essa criança terá grande criatividade ou pensamento profundo. Ela poderia ser uma criança de 5 anos com maturidade de 10 anos, mas com habilidades qualitativamente “normais”. Essa criança se tornaria um adulto com pensamento qualitativamente semelhante ao de outros adultos, exceto por ser mais veloz ou por atingir a maturidade adulta mais cedo, e resolveria problemas normais mais rapidamente e mais facilmente do que os adultos normais, mas não há nenhuma razão para supor que ela resolveria problemas que o adulto normal não resolveria.

Por outro lado, uma criança de 5 anos com idade mental de 5 ou 6 anos, portanto com QI mais baixo que a do nosso primeiro exemplo, poderia ser muito criativa. Como ela seria menos precoce que a outra, poderia não se destacar em testes tradicionais, porque essa criatividade simplesmente não seria detectada nas questões destinadas às crianças de 5, 6 ou 10 anos, nem mesmo nas questões destinadas a adultos, então seu talento passaria despercebido e só se manifestaria quando ela tivesse a oportunidade de encontrar situações nas quais sua criatividade pudesse fazer a diferença. Como resultado, ela se tornaria um adulto genial, com pensamento qualitativamente diferente (e melhor) que o dos adultos ditos “normais”. No primeiro caso, o diagnóstico correto seria “precocidade extrema” e nesse último caso o diagnóstico correto seria “talento” ou, dependendo do nível do talento, o diagnóstico poderia ser “gênio”.

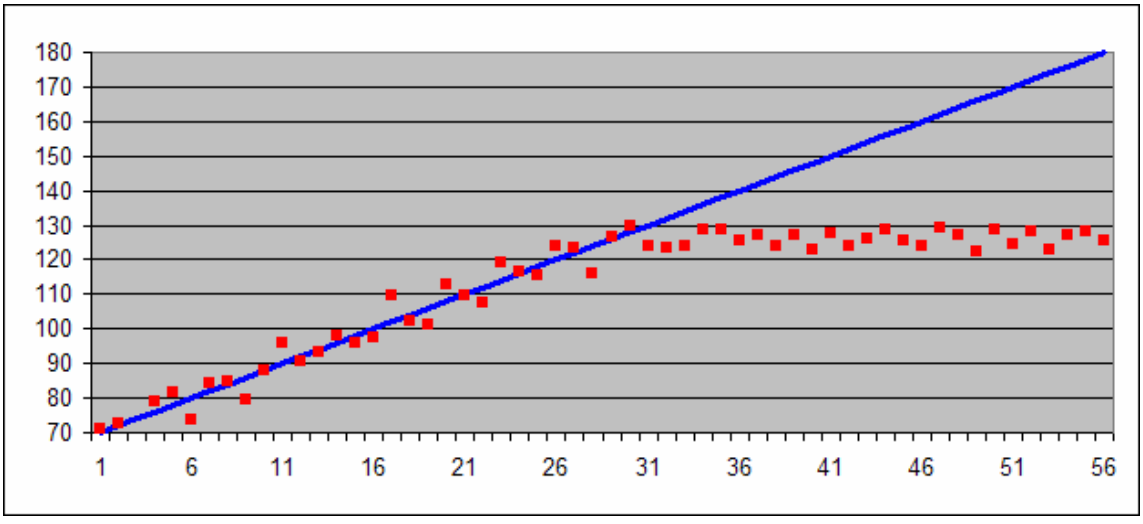
Isso significa que existe um erro em misturar precocidade, talento, genialidade e capacidade de produção intelectual. O que os testes tradicionais podem medir é a precocidade. Eles medem o quão cedo uma criança pode alcançar o nível normal de inteligência. Isso é muito diferente de medir o nível de produção intelectual que a criança terá quando adulta. O que o Sigma Test pretende medir (e há fortes indícios de que efetivamente mede) é a produtividade intelectual, ou, mais precisamente, “a capacidade latente para a produção intelectual”, independente da faixa etária. Por isso é que as 1500 crianças com QIs acima de 135 estudadas por Terman, quando se tornaram adultas, mesmo todas elas somadas não produziram nada comparável à obra de uma única criança com QI 123, como Richard Feynman, que na idade adulta ganhou um prêmio Nobel e foi autor de muitos trabalhos expressivos, sendo considerado um dos maiores gênios do século XX.

O problema não acontece apenas nos casos de crianças. Muito pelo contrário. O problema é ainda mais grave em idade adulta. Um teste como o WAIS ou o Cattell é bom para 95% da população e muito útil para triagens de operários, para aprovação de carteira de habilitação para condutores de veículos, para diagnósticos dentro do intervalo de validade (70 a 130) e para muitas outras aplicações, mas não pode, em hipótese alguma, ser usado para selecionar altos executivos, discriminar entre altos acadêmicos, selecionar para vestibulares concorridos ou para qualquer outra finalidade que envolva cortes acima do percentil 98, porque as funções que essas pessoas terão que assumir geralmente não exigem rapidez na realização de tarefas primárias, que é a principal habilidade medida por esses testes. Em vez disso, essas funções exigem capacidade para formular estratégias complexas e eficientes, para desenvolver e gerir projetos, para administrar simultaneamente grande volume de informações, para tomar decisões em situações realmente difíceis, para conduzir experimentos complexos e toda uma gama de habilidades que nem de longe se assemelham ao que os testes tradicionais medem. O resultado disso é que quando os testes tradicionais são aplicados em situações que extrapolam seus limites de funcionalidade, acabam produzindo resultados totalmente inverossímeis, e os efeitos disso podem ser terríveis, como no já citado caso de Richard Feynman, cujo QI foi medido em 123, mas ele foi seguramente uma das pessoas mais

inteligentes da história e seu verdadeiro QI de raridade estava em torno de 185. Os testes tradicionais podem medir performance mental no intervalo de 70 a 130, cobrindo o espectro indicado abaixo:



Nesse intervalo, o nível de produção intelectual correlata fortemente com a performance mental, por isso se um teste estiver medindo performance mental, o mesmo escore pode ser usado para representar a capacidade de produção intelectual, sem que essa confusão cause distorções significativas. Mas se um teste de performance mental, como o Binet IV, WAIS III ou RAPM III ou versões anteriores, for usado para continuar medindo performance mental em níveis acima de 130, chegando até 187 (1 em 40.000.000), como é o caso do Cattell III, os escores só servirão para representar a performance mental, mas não a capacidade de produção intelectual, porque nos níveis mais elevados a correlação entre performance e potencial é baixa, conforme representa a figura abaixo, em que o eixo x indica a quantidade de acertos (1 a 56), o eixo y indica o QI, a linha azul representa o escore obtido no teste em função da quantidade de acertos e os pontos vermelhos representam as verdadeiras capacidades de produção das pessoas. Conforme podemos observar, os pontos vermelhos se distribuem perto da linha azul no intervalo de QI entre 70 e 130, mas depois disso a linha azul continua subindo e gerando escores cada vez mais altos, porém as capacidades verdadeiras das pessoas que alcançam esses escores continuam no nível de 130.



Isso significa que uma pessoa veloz, mas com pensamento superficial, pode obter escore 180 num teste como o Stanford-Binet IV, enquanto outra pessoa com capacidade de produção

muito maior, porém não tão rápida, pode ter escore de apenas 130 ou 120, como no caso de Feynman. O resultado disso é que os testes convencionais não podem servir para discriminação em altos níveis de habilidade.

Em 1973, para solucionar esse e outros problemas, Kevin Langdon criou o LAIT (*Langdon Adult Intelligence Test*), o primeiro teste de inteligência destinado a medir "corretamente" a capacidade de produção intelectual em altos níveis. Em 1985, Ronald Kent Hoeflin criou o Mega Test e o Titan Test, que durante muitos anos foram os melhores instrumentos psicométricos de alto nível do mundo, inclusive a Mega Society, fundada por Hoeflin em 1982, e que usa o Titan e o Mega como critérios para admissão, foi registrada no Guinness Book de 1990, por ser a sociedade de alto QI mais exclusiva do mundo, com *cut-off* teórico de 1 em 1.000.000, enquanto os testes tradicionais não conseguem discriminar corretamente nada acima de 1 em 100. Os testes de Langdon e Hoeflin foram recebidos com entusiasmo por alguns, e com reservas por outros. As principais objeções a esses testes é que estavam sobrecarregados de cultura especializada, e os matemáticos levavam grande vantagem sobre profissionais de outras áreas. Além disso, os testes estavam em inglês e as questões verbais de associações e analogias exigiam um vocabulário muito extenso na língua inglesa, tornando esses testes quase exclusivamente destinados a anglófonos. Na década de 1990, foram criados muitos novos testes, mantendo os pontos fortes do Mega e do LAIT e tentando eliminar os pontos fracos. Atualmente existem dezenas de testes semelhantes aos de Langdon e Hoeflin, e admite-se que o teto de validade desses instrumentos fica em torno de 4 desvios-padrão acima da média, embora alguns autores reivindiquem para seus testes tetos até 6 desvios-padrão acima da média. Tais reivindicações, no entanto, não encontram respaldo nem no nível de dificuldade das questões nem nos dados estatísticos sobre os escores.

Nesse contexto, o Sigma Test e o Sigma Test VI, criados respectivamente em 1999 e 2001, disponíveis em 16 idiomas (13 e 7, respectivamente), estão conquistando cada vez mais prestígio entre os membros e dirigentes das principais comunidades de alto QI.

Para conhecer as opiniões de alguns proeminentes intelectuais que falam sobre suas impressões a respeito do Sigma Test, basta visitar essa página:

http://www.sigmasociety.com/sigma_opiniao.asp

Portanto a precocidade extrema e a genialidade não estão necessariamente ligadas e os testes tradicionais não servem para diagnosticar corretamente a genialidade. Isso não significa que não existem testes adequados para isso. Os testes existem há pelo menos 30 anos e estão sendo aprimorados constantemente. Porém ainda não começaram a ser usados clinicamente.

Sobre o autor: **Hindemburg Melão Jr.** é detentor de três recordes mundiais em atividades intelectuais, um dos quais está registrado na edição de 1998 do *Guinness Book of Records*, páginas 110-111, é membro honorário em várias associações culturais internacionais, inclusive em Pars Society, na Turquia, para pessoas com QI acima de 180 (em cada 3.500.000 de pessoas, apenas uma tem QI no nível suficiente para ser aprovada), membro honorário em ISI Society, na Inglaterra, para pessoas com QI acima de 151 (em cada 1.400 pessoas, apenas uma tem QI no nível suficiente para ser aprovada), sócio honorário em High IQ Society for Humanity, na Dinamarca, que tem como o objetivo orientar e subsidiar crianças talentosas que vivem em regiões carentes, fundador de Sigma Society, para pessoas superdotadas, que atualmente reúne mais de 200 membros provenientes de 40 países dos 5 continentes, autor do Sigma Test e do Sigma Test VI, disponíveis em 13 e 7 idiomas, respectivamente, publicados em 7 revistas internacionais especializadas em inteligência e testes de QI e em mais de 300 web sites internacionais sobre temas afins. Autor de mais de 350 trabalhos publicados em mais de 120 países, inclusive trabalhos premiados em nível TOP-10 mundial, TOP-19 mundial e TOP-26 mundial.

Mais informações sobre o autor: http://www.sigmasociety.com