

Software para avaliação de aprendizagem utilizando a teoria da resposta ao item

Gilvan Justino, Dalton Francisco de Andrade¹

¹Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC- Brasil
{gilvan,dandrade}@inf.ufsc.br

Abstract. *This paper shows an environment to support the evaluation of the learning process using Item Response Theory, being this last, able to offer information about the difficulty and the discrimination power of the questions, and to evaluate the performance of the student throughout the time, among other resources. It is expected that the environment will allow the educators to make use of some of the rich resources offered by the item response theory.*

Resumo. *Este artigo apresenta um ambiente para apoio ao processo de avaliação de aprendizagem utilizando a Teoria da Resposta ao Item, sendo esta última, capaz de oferecer informações sobre a dificuldade e o poder de discriminação das questões, e avaliar o desempenho do aluno ao longo do tempo, entre outros recursos. Espera-se que o ambiente permita aos educadores disporem de alguns dos ricos recursos oferecidos pela teoria da resposta ao item.*

1. Introdução

A forma mais tradicional para avaliação de desempenho é denominada de Teoria Clássica de Testes (TCT), que se baseia em observar a quantidade de questões corretas dentre um conjunto total de questões, isto é, obter o escore do teste. Uma vantagem desta teoria é que ela é relativamente fácil de interpretar e exige poucas suposições sobre os dados. Por outro lado, a teoria clássica sofre de algumas deficiências ou limitações, de acordo com [Hammer 1999] e [Andrich 1982]:

- a) O escore do estudante não é uma medida absoluta, pois pode variar de teste para teste, dependendo do conteúdo do teste;
- b) É difícil comparar o desempenho de alunos aplicando-se testes diferentes;
- c) A avaliação de desempenho dos estudantes é influenciada pela amostra analisada.

Nas últimas décadas, uma nova teoria, denominada de Teoria da Resposta ao Item (TRI) vem sendo estudada e aplicada com sucesso para construção e análise de testes. A TRI possui várias vantagens em relação à TCT. Seus principais benefícios são conforme [Baker 1992] e [Hambleton 1977]:

- a) Permite construir uma escala para medir o conhecimento dos estudantes, de tal forma que se possa avaliar a proficiência dos estudantes tornando-a independente;

- b) Possibilita obter características das questões (“itens”, no vocabulário da TRI), identificando as questões que realmente contribuem para avaliação do conhecimento;
- c) Permite acompanhar o desenvolvimento de um aluno ao longo do tempo;
- d) Permite comparar resultados de testes aplicados em classes de alunos diferentes;
- e) Permite comparar a dificuldade das questões;

A TRI utiliza um modelo matemático para extrair informações e realizar estimativas das questões e dos estudantes. As estimativas buscam explicar o efeito entre as respostas dos estudantes e seus traços latentes (habilidade/proficiência). O modelo matemático expressa tal relação no formato de equação.

No Brasil, a TRI vem sendo utilizada com sucesso na análise dos dados do Sistema Nacional de Ensino Básico (SAEB), enquanto que no exterior, pode-se citar o *Programme for International Student Assessment* (PISA) [Andrade et al 2000]. Consequentemente pode-se observar a importância no desenvolvimento de ferramentas que suportem a teoria da resposta ao item, auxiliando a análise de testes educacionais.

A teoria da resposta ao item vem sendo utilizada também para resolver problemas de várias outras áreas, como por exemplo, avaliação do nível de satisfação de clientes, qualidade de vida, etc. As ferramentas disponíveis hoje em dia que lidam com a TRI não tratam de um domínio em específico, tornando o seu uso para avaliação de aprendizagem pouco acessível aos avaliadores, pois a TRI utiliza conhecimentos avançados de matemática e estatística. Infelizmente não existem atualmente ferramentas que utilizem a TRI e sejam voltadas para a avaliação de aprendizagem e direcionadas para o uso do avaliador. O objetivo deste trabalho é apresentar um ambiente de avaliação de aprendizagem que utilize a teoria da resposta ao item e que abstraia a complexidade de sua utilização, permitindo assim que o avaliador usufrua de alguns dos recursos da teoria da resposta ao item. Apesar disto, é necessário que o usuário tenha algum conhecimento das ferramentas (recursos) fornecidos pela teoria da resposta ao item.

As próximas seções estão assim organizadas: Na segunda seção é apresentada a teoria da resposta ao item, abordando principalmente os recursos da TRI que foram desenvolvidos neste trabalho. A seção seguinte trata da especificação do ambiente e na última seção são apresentadas algumas conclusões e sugestões para trabalhos futuros.

2. A Teoria da Resposta ao Item

Os estudos sobre a Teoria da Resposta ao Item surgiram na década de 1960 com objetivo de disponibilizar recursos não encontrados na Teoria Clássica de Testes, trazendo informações mais ricas para análise de testes.

A TRI assume a existência de um traço latente ou constructo, isto é, uma habilidade que não pode ser diretamente observável, sendo difícil de mensurá-la. Entretanto, a habilidade pode ser estimada a partir de um modelo matemático que leva em consideração características das questões e a probabilidade do indivíduo acertar uma determinada questão dado o seu grau de habilidade.

O modelo matemático geralmente utilizado é o modelo de Birnbaum [Andrade et al 2000] e [Baker and Kim 2004], também conhecido como função logística de 3 parâmetros que caracteriza cada questão em 3 parâmetros, conforme pode ser visto em (1):

$$P(\phi) = c + (1-c) \frac{1}{1 + e^{-a(\phi-b)}} \quad (1)$$

Onde: b é o parâmetro de dificuldade
 a é o parâmetro de discriminação
 c é o parâmetro de acerto ao acaso
 ϕ é o nível de habilidade/traço latente/proficiência.

O parâmetro b é capaz de informar o grau de dificuldade da questão, enquanto o parâmetro a é utilizado para identificar o poder que a questão tem em discriminar a habilidade dos estudantes. Já o parâmetro c é utilizado para informar a chance que um indivíduo tem em acertar a questão ao acaso.

O modelo de Birnbaum pode ser expresso graficamente através do que se chama de “Curva característica do item” (CCI). Um exemplo pode ser visto na Figura 1:

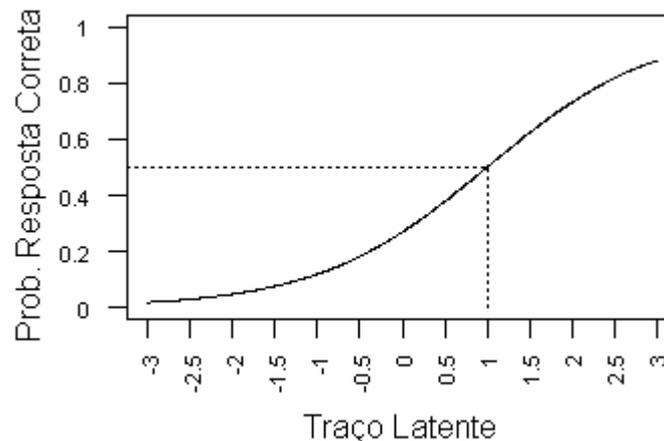


Figura 1 – Curva característica do item

O eixo vertical representa a probabilidade de acertar a questão enquanto o eixo horizontal indica o traço latente dos estudantes, sendo portanto, a escala de conhecimento da proficiência estudada. Para a medida “Traço latente”, quanto mais à direita, isto é, quanto maior o seu valor, maior será a proficiência do estudante. Desta forma, segundo a Figura 1, um indivíduo com traço latente igual a 3 tem probabilidade superior a 80% em acertar tal questão. O inverso também pode ser observado, isto é, um indivíduo com baixo traço latente (-3) tem probabilidade próxima a 0 em acertar a questão.

As propriedades de discriminação e acerto ao acaso também são identificáveis a partir do gráfico. Uma questão com discriminação alta apresenta uma curva mais inclinada, enquanto uma curva menos inclinada indica uma questão com baixa discriminação. O acerto ao acaso é observável a partir do ponto da curva mais próximo ao eixo horizontal. No caso da Figura 1, pode-se observar que a curva está muito próxima do eixo horizontal indicando que esta seria uma questão com baixa probabilidade de acerto ao acaso.

A Curva Característica do Teste é similar à CCI exceto que ela é obtida a partir de todo o conjunto de questões. Na prática, é a soma de probabilidades computadas com a CCI de todos os itens e por isso é denominado de *escore verdadeiro*. Esta medida pode ser utilizada como sendo a nota de um indivíduo num determinado teste.

A Função de Informação do Item é outro recurso da TRI utilizada para indicar o grau de precisão da estimativa da habilidade de um determinado item. É útil para identificar as questões que são realmente relevantes. Já a Função de Informação do Teste faz o mesmo para o teste como um todo.

Outros recursos da TRI envolvem também a capacidade de comparar o desempenho de grupos de estudantes sobre determinados aspectos em relação à mesma série ou séries diferentes. Este processo é denominado de *equalização de teste* [Senno 2006].

Mais informações sobre as ferramentas da TRI e os algoritmos utilizados nos cálculos da teoria da resposta ao item podem ser encontrados em [Andrade et al 2000], [Azevedo 2003], [Baker 1992], [Senno 2006] e [Zubairi 2006].

3. Especificação do software

O ambiente foi elaborado para permitir que o professor pudesse usufruir dos principais recursos da TRI, incluindo a curva característica do item, curva característica do teste, função de informação do item e função de informação do teste. Para realizar a análise através da TRI, o professor deve informar o teste bem como o gabarito das questões. Após aplicado o teste, o professor deve submeter ao software as respostas individuais de cada aluno. Estas são as informações necessárias para se realizar a análise com a TRI. Em seguida, os algoritmos do modelo da TRI podem ser executados para extração de dados. Como saída, o software disponibiliza gráficos e relatórios para análise do professor.

Para implementação do modelo de Birnbaum foi aproveitada boa parte da implementação realizada por [Pinheiro 2006], no seu trabalho de mestrado, que construiu o modelo utilizando a linguagem R. A linguagem R é uma linguagem funcional, pois dá ênfase à aplicação de funções e avaliação de expressões, ao contrário das linguagens de programação tradicionais. Esta característica permite a construção de algoritmos utilizando-se operações mais intuitivas que as utilizadas pelas linguagens convencionais, sendo portanto um fator importante para desenvolvimento dos complexos algoritmos matemáticos e estatísticos exigidos pela teoria da resposta ao item. Informações sobre a linguagem R podem ser obtidas em [Venables 2003].

R não é uma linguagem usada para escrever interfaces gráficas (GUI), apesar de possuir o pacote tcl-tk, que é utilizado para escrever algumas interfaces baseadas em janelas. O principal objetivo de R é permitir a implementação de algoritmos para processar dados. Por isso foi utilizada a linguagem Delphi para construção do ambiente gráfico.

Para que um programa R seja executado a partir de uma outra aplicação escrita em outra linguagem, é necessário fazer uso de algum pacote de software para estabelecer uma comunicação com o Engine R. Na literatura, citam-se os pacotes RD-COM e SJava. Para o software desenvolvido foi utilizado RD-COM.

A Figura 2 exibe a organização dos componentes utilizados na construção do software.

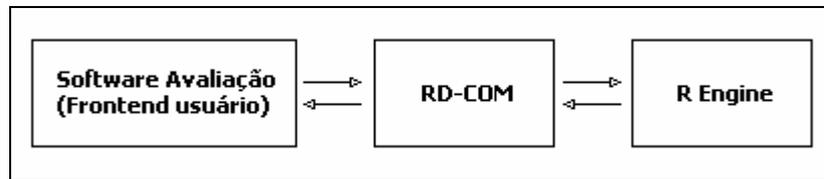


Figura 2 – Componentes utilizados para o projeto de avaliação através da TRI

O frontend do usuário (GUI) interage com R através do RD-COM. Como os algoritmos da TRI são construídos integralmente em R, todas as operações de processamento necessariamente devem ser solicitadas ao Engine R através do RD-COM. A interface gráfica desenvolvida em Delphi se encarrega de solicitar as informações ao usuário e conduzir os dados para o Engine R, além de exibir as saídas geradas pelos algoritmos.

A Figura 3 exibe exemplos de chamadas ao engine R, a partir do código Delphi. No contexto apresentado, RCon é a classe que permite a conexão com o engine R através de COM, enquanto o método ENR encarrega-se de submeter uma instrução para execução.

```

RCon.ENR('items.deal <- dataset.info[3,1:(ncol(dataset.info)-1)]');
RCon.ENR('items.names <- dataset.original[4,2:ncol(dataset.original)]');
RCon.ENR('items.count <- ncol(dataset.info)-1');
  
```

Figura 3 - exemplo

A Figura 4 apresenta os principais elementos da especificação em UML para o desenvolvimento deste software.

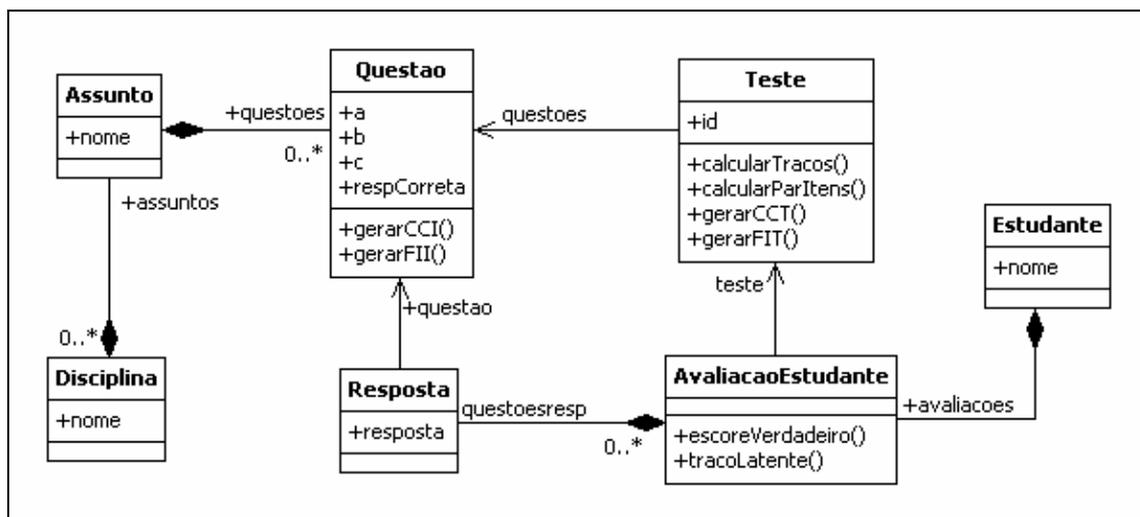


Figura 4 – Diagrama UML do software

As classes Assunto, Disciplina e Questão constituem o banco de itens, isto é, um cadastro de questões catalogadas para serem aplicadas nos testes. Interfaces para manutenção do banco de itens estão disponíveis no software. A classe Questao armazena as propriedades do item segundo a teoria da resposta ao item, isto é, mantém

os parâmetros a , b e c , além da resposta correta. O método gerarCCI é utilizado para processar o gráfico da curva característica do item, enquanto que o método gerarFII é utilizado para processar o gráfico da função de informação do item. Ambas as funções foram construídas em R. Estes métodos somente podem ser executados depois que o cálculo de habilidades e parâmetros de itens forem executados.

A classe Teste possui uma associação denominada *questoes* que é útil para referenciar as questões do banco de itens que foram ou serão aplicadas num determinado teste. Os métodos calcularParItens e calcularTracos são utilizados respectivamente para calcular os parâmetros dos itens e calcular os traços latentes dos estudantes. Os traços latentes dos estudantes somente podem ser calculados caso os parâmetros dos itens forem conhecidos. Sempre que uma questão é utilizada num teste e as estimativas são calculadas, o software armazena os parâmetros encontrados para serem utilizados em futuros testes.

Os métodos gerarCTT e gerarFIT correspondem respectivamente às funções que geram a curva característica do teste e a função de informação do teste. Estes métodos somente podem ser solicitados assim que os cálculos dos parâmetros dos itens e habilidades tiverem sido executados previamente. Ambas as funções desviam o fluxo de execução para programas R processarem a geração do gráfico apropriado.

A figura seguinte (Figura 5) apresenta uma das saídas fornecidas pelo software.

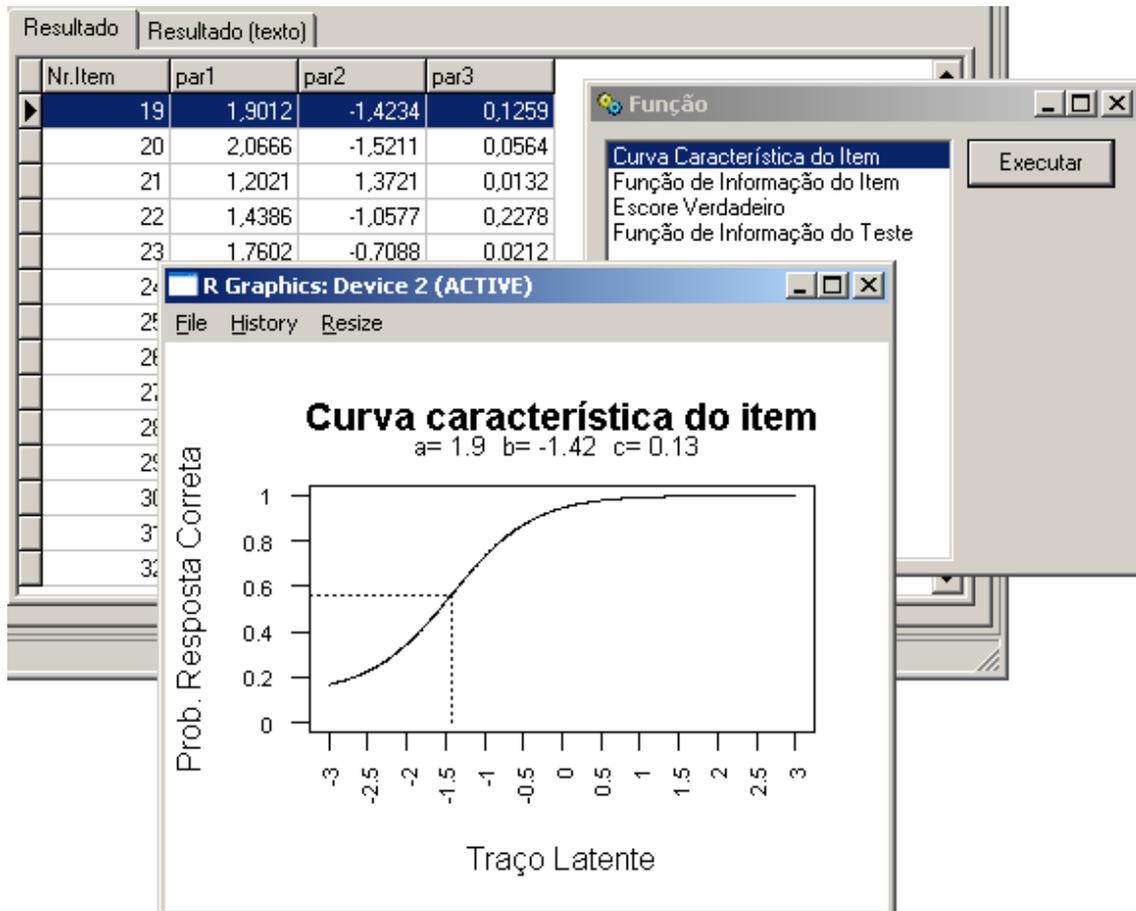


Figura 5 – Ambiente de avaliação usando TRI

Neste caso, os parâmetros dos itens foram calculados e exibidos na tela. A questão de número 19 foi destacada para visualização da sua curva característica. Pelo que pode ser observado, esta questão tem nível de dificuldade fácil pois estudantes com proficiência baixa têm boa probabilidade de acertá-la. Além disso, pode-se concluir que os estudantes têm cerca de 13% de chance em acertá-la ao acaso. Através do formato da curva, pode-se concluir também que a curva tem boa discriminação.

A análise da curva característica do item permite ao avaliador decidir se a questão é realmente útil para ser aplicada para avaliar a proficiência dos estudantes.

As ferramentas atualmente desenvolvidas e descritas até aqui constituem apenas alguns dos recursos que a TRI oferece. Várias outras ferramentas podem ser construídas. O ambiente foi modelado para suportar plug-ins escritos em R e segue a especificação apresentada em [Justino 2007]. O plug-in tem acesso a uma série de informações fornecidos pelo software desenvolvido. Tais informações estão descritas na tabela 1.

Tabela 1 – Estruturas fornecidas para construção de plug-in

| Objeto R | Tipo de dado | Descrição |
|--------------------|--------------|---|
| Dataset | Matriz | Identificação dos estudantes com suas respectivas respostas |
| dataset.info | Matriz | Informações sobre o gabarito e questões. |
| items.rightanswers | Vetor | Gabarito |
| items.numoptions | Vetor | Número de opções para cada questão |
| Items.estimates | Vetor | Parâmetro dos itens |
| items.name | Vetor | Identificação das questões |
| dataset.answers | Vetor | Respostas dos estudantes |

O plug-in deve ser mantido em formato compactado (zip) e deve conter, além dos programas escritos em R, um arquivo denominado index.xml, que publica as funções e propriedades da ferramenta desenvolvida.

4. Conclusões

O uso da linguagem R permitiu simplificar a construção do modelo pois se trata de uma linguagem apropriada para implementação dos algoritmos utilizados pela teoria da resposta ao item e que requerem a execução de procedimentos matemáticos e estatísticos avançados.

O frontend (GUI) facilita a submissão do teste para análise através da TRI pois o avaliador não precisa acessar a shell/console do R para execução dos comandos.

A construção do software através de plug-ins permite que outros desenvolvedores ampliem as funcionalidades do software, estendendo as suas funcionalidades originais.

A expectativa deste trabalho é que o ambiente possa contribuir com a disseminação do uso da teoria da resposta ao item na avaliação de desempenho de estudantes. Vale lembrar que apesar do projeto prever a simplificação do uso da teoria, ainda assim se faz necessário que o avaliador faça um estudo da teoria da resposta ao item a fim de compreender suas ferramentas.

Como atividades futuras, pretende-se construir os plugins para implementar a equalização de testes e os métodos apresentados por [Azevedo 2003] e [Tavares 2006]. Além disso, pretende-se utilizar o software desenvolvido na unidade de Ensino Técnico Senai da cidade de Blumenau a fim de agrupar os alunos em turmas com conhecimento mais homogêneo entre si.

Referências

- Andrade, Dalton Francisco de and Tavares, Heliton Ribeiro and Valle, Raquel da Cunha (2000) “Teoria da Resposta ao Item: Conceitos e Aplicações”. Associação Brasileira de Estatística – ABE.
- Andrich, D. (1982) “An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern”, In *Education Research and Perspectives*. p. 9, 95-104.
- Azevedo, Caio Lucidius Naberezny (2003) “Métodos de estimação na teoria de resposta ao item”. Dissertação de Mestrado. Instituto de Matemática e Estatística, USP.
- Baker, Frank B. (1992) “Item Response Theory: Parameter Estimation Techniques”, Marcel Dekker, Inc.
- Baker, Frank B and Kim, Seock-Ho (2004) “Item Response Theory – Parameter Estimation Techniques”, 2nd edition. Marcel Dekker, Inc.
- Hambleton, K. Ronald (1977) Latent Traits Models and Their Use in the Analysis of Education Test Data. In *Journal of Educational Measurement*, Vol. 14. Nr. 2, Applications of Latent Trait Models (Summer, 1977). p. 75-96.
- Hammer, Robert J. (1999) “Item Response Theory”, In *The Counseling Psychologist*, vol. 27. p 353-383.
- Justino, Gilvan (2007) e Andrade, Dalton Francisco de “An environment specification for item response theory”, In *Third International Conference on Intelligent Computing and Information Systems*.
- Pinheiro, Conrad Elber (2006) “Implementação de métodos estatísticos para avaliação educacional no software R”. Dissertação de Mestrado. IME, USP.
- Senno, Rosangela Molini (2006) “Métodos de Equalização na Teoria Clássica e na Teoria da Resposta ao Item”. Dissertação de Mestrado. IME, USP.
- Tavares, Heliton Ribeiro e Andrade, Dalton Francisco de (2006) “Item Response Theory for Longitudinal Data: Item and Population Parameters Estimation Test”, In *Test*, V. 15, n.1, p. 97-123.
- Venables, W. N. e Smith, M. (2003) “An introduction to R”, <http://cran.r-project.org/doc/manuals/R-intro.pdf>, Junho/2006.
- Zubairi, Ainol Madziah e Kassim, Noor Lide Abu (2006) “Classical and Rasch Analyses of Dichotomously Scored”. In *Journal of the Malaysian English Language Teaching Association (MELTA)*, vol. 2