UNIVERSIDADE FEDERAL DA BAHIA

FACULDADE DE EDUCAÇÃO

NÚCLEO DE PÓS-GRADUAÇÀO EM EDUCAÇÃO - NPGE

LYS MARIA VINHAES DANTAS

AS CONTRIBUIÇÕES DAS POLÍTICAS DE AVALIAÇÃO EDUCACIONAL EM LARGA ESCALA: O CASO DA AVALIAÇÃO DE APRENDIZAGEM NA BAHIA.

Salvador

2009

LYS MARIA VINHAES DANTAS

AS CONTRIBUIÇÕES DAS POLÍTICAS DE AVALIAÇÃO EDUCACIONAL EM LARGA ESCALA: O CASO DA AVALIAÇÃO DE APRENDIZAGEM NA BAHIA.

Tese apresentada ao Programa de Pesquisa e Pós-graduação em Educação, Faculdade de Educação, Universidade Federal da Bahia, como requisito para obtenção do grau de doutor em educação.

Orientador: Prof. Dr. Robert E. Verhine.

Salvador

2009

UFBA / Faculdade de Educação – Biblioteca Anísio Teixeira

D192 Dantas, Lys Maria Vinhaes.
As contribuições das políticas de avaliação educacional em larga escala : o caso da avaliação de aprendizagem na Bahia / Lys Maria Vinhaes Dantas. – 2009. 258 f. : il.

Orientador: Prof. Dr. Robert E. Verhine.
Tese (doutorado) – Universidade Federal da Bahia. Faculdade de Educação, 2009.

1. Avaliação educacional - Bahia. 2. Políticas públicas. 3. Escolas públicas – Avaliação. I. Verhine, Robert E. II. Universidade Federal da Bahia. Faculdade de Educação. III. Título.
.

CDD 371.26098142 – 22. ed.

LYS MARIA VINHAES DANTAS

AS CONTRIBUIÇÕES DAS POLÍTICAS DE AVALIAÇÃO EDUCACIONAL EM LARGA ESCALA: O CASO DA AVALIAÇÃO DE APRENDIZAGEM NA BAHIA.

Tese apresentada ao Programa de Pesquisa e Pós-graduação em Educação, Faculdade de Educação, Universidade Federal da Bahia, como requisito para obtenção do grau de doutor em educação.

Aprovada em 29 de julho de 2009.

Banca examinadora:

Robert Evan Verhine – Orientador
Doutor em Educação pelo Universitat Hamburg, Alemanha
Universidade Federal da Bahia

Romualdo Luiz Portela de Oliveira
Doutor em Educação pela Universidade de São Paulo
Pós-Doutor pela Cornell University, Estados Unidos
Livre Docente pela Universidade de São Paulo
Universidade de São Paulo

Kátia de Siqueira Freitas
Doutor em Educational Administration pela The Pennsylvania State University, Estados Unidos
Pós-Doutor pela The Pennsylvania State University, Estados Unidos
Universidade Católica do Salvador

José Albertino Carvalho Lôrdelo
Doutor em Educação pela Universidade Federal da Bahia
Universidade Federal da Bahia

Dora Leal Rosa
Doutor em Educação pela Universidade Federal da Bahia
Universidade Federal da Bahia

[…] no evaluator assumes the role of being in possession of any indisputable "truth" that must be accepted by others. (REBOLLOSO *et alii*, 2002:14)

Resumo

DANTAS, Lys Maria Vinhaes. As contribuições das políticas de avaliação educacional em larga escala: o caso da avaliação de aprendizagem na Bahia. 259 f. il.2009. Tese (Doutorado) – Faculdade de Educação, Universidade Federal da Bahia, Salvador, 2009

Esse estudo investigou as contribuições de políticas de avaliação educacional, quando implementadas em larga escala, para as escolas públicas. Escolheu como foco a política de Avaliação da Aprendizagem (AA), implementada pelo Governo do Estado da Bahia junto às redes estadual e municipais (parceiras do Estado) no período de 1999 a 2004 (com primeira aplicação de provas em 2001). Entendeu como contribuições o conjunto de elementos de Utilidade e Uso. Para tanto, adaptou os sete indicadores da categoria Utilidade do modelo de meta-avaliação do JCSEE e do *checklist* decorrente elaborado por Stufflebeam (1999) e incluiu, no modelo adaptado, um indicador para a percepção de utilidade (U8). Dentre os 66 itens verificadores dos sete indicadores da categoria Utilidade (U1 a U7), 44 foram observados na AA (67%), o que apontava para uma probabilidade média a alta de concretização do uso, reforçada por uma percepção de utilidade da AA relatada pelas escolas (U8). A partir da discussão sobre usos instrumentais e conceituais, este trabalho propôs uma segunda categoria de análise: Uso. Inicialmente, focalizou o uso dos resultados da avaliação para a tomada de decisões (uso instrumental clássico), buscando relatos de usos feitos pelas escolas públicas, encaminhados à equipe central da avaliação por meio dos Relatórios do Diretor (RD). Expandiu a consulta sobre usos para além dos resultados, incluindo outros elementos da avaliação. Estes se mostraram mais freqüentes, como o uso das matrizes de referências. Por essa razão, advoga-se aqui a ampliação da noção de uso instrumental para além dos resultados. Em seguida, voltou-se para o atingimento dos objetivos da política, na busca por uma relação entre o uso e a finalidade da política, lançando mão das variações nas taxas oficiais (aprovação, reprovação, abandono e proficiência em português e em matemática). Por fim, o estudo buscou relatos de usos conceituais, fundamentais para o entendimento de como uma política de avaliação pode afetar seus *stakeholders*. Em especial, os resultados da pesquisa mostraram o efeito da AA para o compartilhamento de uma visão sobre a realidade de ensino, sendo este o uso conceitual mais relatado, de maneira positiva, pelas escolas. Respeitados os limites da não representatividade dos respondentes e de uma possível ritualização nas respostas das escolas, os achados apontaram para um uso real da avaliação, com efeito no desempenho do alunado da 4ª série tanto em Português quanto em Matemática. Para finalizar, foi feita uma reflexão sobre a relação entre uso e o atingimento da finalidade da política de avaliação. Diferente do que possa parecer, tal relação não é direta ou linear. Propõe-se aqui que a discussão a ser feita, em lugar de estar focada sobre uso x não uso, deveria ser concentrada em análises sobre se os usos feitos contribuem ou não para a melhoria da qualidade da Educação.

Palavras-chave: avaliação educacional; políticas públicas; utilidade da avaliação; uso da avaliação.

Abstract

DANTAS, Lys Maria Vinhaes. Educational evaluation contributions: the case of the Student Assessment Project in Bahia. 259 f. il.2009. Doctoral Thesis – College of Education , Federal University of Bahia, Salvador, 2009

The present study dealt with the contributions of educational evaluation public policies to elementary public schools. It focused on the Student Assessment Project (AA) implemented by the State of Bahia-Brazil during the 1999-2004 period (first tests administered in 2001). In the context of this investigation, contributions were comprised of elements of evaluation utility and evaluation use. At first, this study adapted the Utility category of the meta-evaluation model proposed by the JCSEE and the meta-evaluation checklist created by Stufflebeam (1999) and added a new standard, the stakeholder's perception of utility. Among the 66 final checkpoints utilized, 44 were observed in the AA experience, pointing to a medium to high probability of use, reinforced by a positive perception of utility the schools manifested. The category Use was proposed based on the discussion of types of use, mainly instrumental and conceptual, found in the evaluation literature. This research investigated the use of the evaluation results for decision making reported by the schools. It also identified the use of other evaluation elements, such as test specifications, and found these elements were more frequently used than the results. The next step was to examine the relation between the evaluation use and the fulfillment of the overall goal of the evaluation policy. The last phase concentrated on the conceptual use reported by the schools. The research results showed that the AA was positively related to affecting the way teachers and principals shared a vision of their work and teaching environment. Regardless of the non representativeness of the sample involved in the research and of a possible ritual school behavior, the findings indicate better 4^{th} graders performance both in Portuguese and Math in the group of schools involved with the AA, but these results were not necessarily related to the intended uses of the policy. Besides discussing the use of the JCSEE Utility category, this study presents two main contributions: it reflects on the concept of instrumental use and discusses that the evaluation use does not necessarily leads to the achievement of the public policies goals.

Key-words: educational evaluation; public policies; meta-evaluation; evaluation utility; evaluation use.

Lista de Abreviaturas e Siglas

AA	Política de Avaliação de Aprendizagem
ABAVE	Associação Brasileira de Avaliação Educacional
ACE	Avaliação das Condições de Ensino
AD	Política de Avaliação de Desempenho
AEA	*American Evaluation Association*
AERA	*American Educational Research Association*
AIR	*American Institutes for Research*
ANRESC	Avaliação Nacional do Rendimento no Ensino Escolar (Prova Brasil)
ANSI	*American National Standards Institute* (ANSI).
APA	*American Psychological Association*
CBA	Ciclo Básico de Aprendizagem
CEE	Conselho Estadual de Educação
COPE	Coordenação de Projetos Especiais
CPA	Comissão Própria de Avaliação
DIREC	Diretoria Regional da Secretaria da Educação da Bahia
DOE	Diário Oficial do Estado
ENADE	Exame Nacional de Avaliação e Desenvolvimento dos Estudantes
ENC	Exame Nacional de Curso (Provão)
ENEM	Exame Nacional do Ensino Médio
EUA	Estados Unidos da América
FHC	Fernando Henrique Cardoso
FAPEX	Fundação de Apoio à Pesquisa e à Extensão
FUNDEB	Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação
FUNDEF	Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério
GPBT	Gestão Pública para um Brasil de Todos
GP	Gestão Pública
IDEB	Índice de Desenvolvimento da Educação Básica
IDH	Índice de Desenvolvimento Humano
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais "Anísio Teixeira"
ISP	Centro de Estudos Interdisciplinares para o Setor Público
JCSEE	*Joint Committee on Standards for Educational Evaluation*
LDB	Lei de Diretrizes e Bases
MEC	Ministério da Educação
NAEP	*National Assessment of Educational Progress*
NCME	*National Council on Measurement in Education*
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
ONU	Organização das Nações Unidas
PCN	Parâmetros Curriculares Nacionais
PDE	Plano de Desenvolvimento da Educação
PDE	Plano de Desenvolvimento da Escola
PISA	Programa Internacional de Avaliação de Estudantes
PNE	Plano Nacional da Educação
PNUD	Programa das Nações Unidas para o Desenvolvimento
PREAL	*Programa de Promoción de la Reforma Educativa en América Latina y el Caribe*
RD	Relatório do Diretor

SAEB	Sistema Nacional de Avaliação da Educação Básica
SAEPE	Sistema de Avaliação do Estado de Pernambuco
SEC	Secretaria da Educação do Estado da Bahia
SIMAVE	Sistema Mineiro de Avaliação da Aprendizagem
SINAES	Sistema Nacional de Avaliação e Progresso do Ensino Superior
SPDE	Superintendência de Políticas e Diretrizes Educacionais
SUDEB	Superintendência de Desenvolvimento da Educação Básica
SUPAM	Superintendência de Articulação Municipal
SUPAV	Superintendência de Acompanhamento e Avaliação do Sistema Educacional
SUPEC	Superintendência da Gestão Escolar
SUPEN	Superintendência de Ensino
TIMMS	*Third International Mathematics and Science Study*
UFBA	Universidade Federal da Bahia
UNEG	*United Nations Evaluation Group*
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura
UNICEF	*United Nations Children´s Fund*

Lista de Ilustrações

Lista de Quadros

Lista de Tabelas

Sumário

1. Introdução[1]

Desde os anos 90, o Brasil tem sido palco para a formulação e a implementação de políticas de avaliação e, na área da Educação, de avaliações realizadas em larga escala. O sucesso do SAEB – Sistema Nacional de Avaliação da Educação Básica –, lançado em 1990, contribuiu muito para o desenvolvimento de avaliações próprias dos estados brasileiros e para a proposta de novos tipos de abordagem federal, mais recentes, como a ANRESC (Avaliação Nacional do Rendimento no Ensino Escolar, conhecida como Prova Brasil, com primeira aplicação em 2005) e a Provinha Brasil, essa última com o objetivo de avaliar a alfabetização no país (primeiro teste aplicado em março de 2008). Em 1998, o Ensino Médio recebeu também o seu mecanismo de avaliação, com crescente participação dos estudantes no ENEM (Exame Nacional do Ensino Médio) desde então. Na Educação Superior, em 1995 os cursos e instituições passaram a ser avaliados por três eixos de ações (Censo de Educação Superior, Avaliação das Condições de Ensino – ACE e o Exame Nacional de Cursos - ENC), com predominância do ENC (conhecido como Provão), e hoje contam com o SINAES (Sistema Nacional de Avaliação da Educação Superior, implementado no início de 2004).

Esse *boom* foi bastante influenciado pelo movimento internacional de Educação para Todos, com o qual o país assina um compromisso (ver *Declaração Mundial sobre Educação para Todos, Jomtiem, Tailândia, 1990*[2]), e com os movimentos de reforma do estado, diante do cenário de recursos escassos com que se defrontava e defronta o setor público. Pressionando os Estados estava uma sociedade civil que começava a se organizar e que passou a ser muito mais exigente em termos de cobrança de qualidade no serviço público, por um lado, e de responsabilização nas ações públicas, por outro. Uma resposta dada por vários países a este cenário foi a descentralização das ações e a concessão de uma maior autonomia para os implementadores das políticas públicas. O Brasil seguiu esta tendência.

Em tal panorama a avaliação se fortaleceu como a contrapartida da autonomia, como uma forma de prestação de contas do Estado e de imputabilidade dos gestores (*accountability*). Contribuiu para o fortalecimento o fato de que a divulgação dos resultados das avaliações, obtidos de maneira rigorosa e sistemática, atendia aos anseios da sociedade civil por maior transparência. Em

1 Trabalho parcialmente financiado pela CAPES por meio de bolsa de doutorado entre os anos 2005-2007.
2 Texto na íntegra pode ser lido no endereço http://unesdoc.unesco.org/images/0008/000862/086291por.pdf

decorrência, a partir dos anos 90, o foco da avaliação muda. De maneira geral, os Estados em processo de descentralização passariam a se preocupar menos com os processos e mais com os seus resultados.

Tudo isso teve e tem um alto custo. As políticas de avaliação competem com as demais políticas por fundos sempre insuficientes e, em um determinado programa, a avaliação é frequentemente acusada de usar os recursos que poderiam estar sendo aplicados na própria intervenção. Para dar um exemplo desse custo, segundo dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP, a primeira aplicação da Prova Brasil custou R$ 54.926.284,68 (cinqüenta e quatro milhões, novecentos e vinte e seis mil, duzentos e oitenta e quatro reais e sessenta e oito centavos)[3] ao Governo Federal (INEP, 2006).

Além disso, talvez por ser a cultura da avaliação relativamente nova no Brasil, as mudanças de governo (em qualquer das esferas) têm, em vários casos, resultado na descontinuidade ou na alteração de programas de avaliação existentes, com perda de séries históricas de dados e de possibilidade de comparação de resultados ao longo dos anos. Tais mudanças raramente têm sido embasadas nos resultados de pesquisas sobre as avaliações. Um exemplo: Souza (2003: 180), ao refletir sobre os usos do SAEB, restringiu-os ao "fortalecimento do papel regulador do Estado, por meio da responsabilização das unidades federadas pelos resultados escolares" e afirmou que o delineamento "não tem potencial para produzir alterações nas práticas escolares, de ensino e de aprendizagem, no sentido de seu aprimoramento". É esse SAEB que, em 2005, passou a ser aplicado censitariamente a todas as escolas, via Prova Brasil. Seria apenas o fato de uma aplicação amostral, em espiral, o elemento que definia a não potencialidade do SAEB em impactar a escola argumentada por Souza? As modificações porque passam as avaliações, como o SAEB, têm sido mais influenciadas pelos posicionamentos dos governantes que sobre um quadro teórico ou sobre uma análise empírica que favoreça a decisão.

Além de serem caras e poderem sofrer modificações drásticas sem uma base de informações que favoreça tais decisões, há uma grande preocupação com a (não) utilização dos resultados das avaliações. Como posto por Helene,

> A grande utilidade de um sistema de avaliação é permitir o estabelecimento de políticas que venham a corrigir os problemas detectados. Entretanto, e infelizmente, não é esse o caso do Brasil. Parece que, como um Narciso às

3 A dotação orçamentária final do Inep, no exercício de 2005, foi de R$ 267.654.418,00 (duzentos e sessenta e sete milhões, seiscentos e cinqüenta e quatro mil e quatrocentos e dezoito reais), de acordo com dados do Relatório de Gestão do INEP para o ano de 2005.

> avessas, ficaremos a contemplar a feiúra de nosso sistema educacional, sem intervir, até sermos inteiramente consumidos (HELENE, s/d: 12).

A preocupação com a não utilização dos dados é tão antiga quanto o são as abordagens avaliativas e mais recentemente voltou ao centro das atenções, com a divulgação do IDEB (Índice de Desenvolvimento da Educação Básica). Em que pese a mudança positiva dos resultados nas séries avaliadas, estudiosos de Educação do País (como Iza Locatelli, Maria Helena G. Castro e João Batista Oliveira) continuam preocupados com o fato de que ela é pequena, além de não necessariamente informar o que foi melhorado. Ademais, como posto por Ravela e outros (2008), os ministérios de educação têm demonstrado capacidade limitada para responder aos problemas identificados pelas avaliações; para esses autores, falta uma maior articulação entre avaliação, desenvolvimento curricular, formação inicial e desenvolvimento profissional dos docentes.

Os problemas das avaliações estão longe de serem sanados, mas a demanda por informações consistentes e precisas é maior a cada dia. A solução, portanto, não passa por eliminar as políticas de avaliação e sim por melhorá-las, como discutido por Verhine (2008) e Schwartzman (2005).

> Não é que as avaliações consigam atender a todas as questões que preocupam, mas os problemas associados às avaliações parecem ser claramente preferíveis aos problemas associados a sua não existência, como a falta de parâmetros, a ausência de critérios, os desperdícios e injustiças na distribuição de recursos, e a impossibilidade de estabelecer políticas consistentes para sistemas de educação de massas. Dos problemas existentes, muitos são de natureza técnica, e podem ir sendo superados na medida em que os estudos, as pesquisas e as metodologias de avaliação se desenvolvam. (SCHWARTZMAN, 2005:31)

A busca por melhoramentos nas avaliações tem, em muitos casos, acompanhado a procura por refinamentos em pesquisa. Desde a década de 50, pesquisadores e associações têm sugerido padrões de qualidade para a pesquisa científica, com foco inicial nos instrumentos e coleta de dados. Em 1969, Scriven propõe o termo meta-avaliação para designar a avaliação de uma avaliação e, nessa introdução, argumenta que pesquisa científica e avaliação, embora tenham muitas semelhanças, são diferentes e precisam ser tratadas diferentemente (SCRIVEN, 1969).

Cinco anos depois, em 1974, Stufflebeam publica um documento com dimensões e critérios de meta-avaliação (a partir de um trabalho anterior desenvolvido com Guba). Três associações (*American Psychological Association*, *American Educational Research Association* e *National Council of Measurement in Education*) se juntam, formando o *Joint Committee on Standards for Educational Evaluation* (JCSEE), e se movimentam no sentido de determinar os padrões de qualidade da avaliação acrescentando a dimensão Utilidade à proposta original de Stufflebeam. Em 1994, o JCSEE lança uma declaração dos critérios de qualidade em avaliação que viria a

influenciar uma série de associações de avaliação em todo o mundo, seja diretamente como no caso de Suíça, Alemanha e Associação Africana ou indiretamente, como na França e Inglaterra (para exemplo, ver as declarações de padrões de qualidade da *Societé Française de L´évaluation*[4], da *United Kingdom Evaluation Society*[5], da *Swiss Evaluation Society*[6] e da *African Evaluation Association*[7], dentre tantos). Além de contribuir para o refinamento das avaliações, o cumprimento dos padrões de qualidade para as avaliações e/ou avaliadores favorece sua credibilidade (HARTZ, 2006) e a identificação de vieses avaliativos (LETICHEVSKY *et alii*, 2007), em sua maioria prejudiciais à obtenção da "boa" informação[8].

O presente trabalho parte de uma discussão sobre o conceito de meta-avaliação, os *standards* de qualidade da avaliação publicados pelo JCSEE em 1994 e um *checklist* de Stufflebeam em 1999 (ainda que modificado), para analisar as contribuições de políticas de avaliação em larga escala, reunindo os elementos preditores do JCSEE e os relatos de usos feitos pelas escolas envolvidas em uma política de avaliação. É a utilização que justifica a avaliação. O esforço avaliativo perde-se inteiramente se seus produtos e processos não são aproveitados pelos *stakeholders*. Alguns autores, dentre os quais Penna Firme e Letichevsky (2002) e Patton (2005), defendem inclusive que a utilidade seja a dimensão definidora de qualidade final da avaliação. Em outras palavras, apenas é considerada uma avaliação de qualidade – e, portanto, justificável – aquela na qual o delineamento, de maneira ética, foi traçado para atender às demandas dos *stakeholders*, resultou em uma implementação técnica rigorosa que possibilitou a coleta da "boa" informação, foi viável e eficiente e, acima de tudo, foi utilizada.

O JCSEE agrupou seus *standards* de qualidade em quatro categorias. De maneira simplificada, as duas primeiras categorias do JCSEE, Precisão e Viabilidade, estão ligadas ao mérito da avaliação e as duas últimas, Propriedade e Utilidade, a seu valor. Mérito e valor serão tratados mais adiante, mas, brevemente, pode-se dizer de mérito como uma condição de qualidade intrínseca da avaliação e de valor como condição de sua contribuição para o bem maior. Uma avaliação de

---

4 *Societé Française de L´évaluation*, fundada em 1999, com foco na avaliação de políticas públicas. Sua "*Charte Française de evaluation de politique et programme public*" está disponível em http://www.sfe.asso.fr/fr/charte-sfe.html.

5 *United Kingdom Evaluation Society*, fundada em 1994, tem foco na promoção e melhoria da teoria, conhecimento da prática e uso da avaliação. Seu Guia de Boas Práticas está disponível em http://www.evaluation.org.uk/Pub_library/Good_Practice.htm

6 *Swiss Evaluation Society* (SEVAL), fundada em 1996, com objetivo de incentivar a troca de informações e experiência no campo da avaliação entre a política, a administração, a academia, as ONGs e o setor privado. Seu *SEVAL Standards* está disponível em http://www.seval.ch/en/ueberuns/index.cfm

7 A *African Evaluation Society* foi fundada em 1999 e tem como objetivo promover e fortalecer a avaliação no continente Africano. Seu *African Evaluation Guidelines* 2000 está disponível em http://www.afrea.org/content/index.cfm?navID=5&itemID=204

grande mérito não necessariamente tem valor, mas, para ter valor, é fundamental que a avaliação seja meritosa. Portanto, para que, em última análise, a Utilidade seja a dimensão definidora da qualidade, é necessário antes que as demais dimensões tenham passado por crivos de qualidade.

Os usos e utilidades, em uma avaliação educacional, têm várias facetas e vêm sendo estudados a partir de cada uma delas, muito influenciadas pela origem teórica dos estudiosos e pesquisadores. Dentre os aspectos de análise de uma avaliação útil, podem-se identificar as necessidades dos *stakeholders*, ou aprofundar o entendimento do quê, na avaliação, é passível de uso, ou o tipo de uso que é ou foi feito, ou ainda buscar os fatores que impactam esse uso, internos ao delineamento avaliativo ou externos, oriundos do contexto. Ao longo do tempo, o foco de interesse mudou do uso instrumental dos resultados da avaliação para as análises de contexto e para as influências que uma política de avaliação possa ter nos diversos níveis de uso – micro, meso ou macro. Há ainda a considerar a percepção de utilidade que os *stakeholders* desenvolvem em relação ao processo avaliativo, que independe das finalidades da avaliação e dos usos concretizados.

A categoria Utilidade do JCSEE aborda apenas uma parte das facetas da avaliação útil, concentrando-se nos aspectos que possam levar a uma utilização da avaliação pelos *stakeholders*. No presente trabalho, essa categoria é empregada para identificar elementos facilitadores de uso no delineamento e implementação de uma determinada política de avaliação, complementada pela percepção de utilidade de *stakeholders* principais, mas é proposta uma segunda categoria de análise: Uso. Essa nova categoria engloba os tipos de uso concretizados pelos mesmos *stakeholders*. Utilidade e Uso da avaliação são tratados como contribuições da avaliação em larga escala e compõem a base do quadro teórico que suporta a análise da pesquisa em tela.

Com a hipótese de que, **em políticas de avaliação educacional em larga escala, os resultados são elementos pouco utilizados e é o acontecimento da avaliação que afeta as instituições em nível micro** (escolas), o presente trabalho teve como foco a política de Avaliação da Aprendizagem, formulada e implementada pelo Estado da Bahia no período de 2001 a 2004, e pergunta:

Quais as contribuições de um ciclo completo da política de Avaliação da Aprendizagem para as escolas por ela envolvidas?

---

8 Segundo Stufflebeam (1974:5), "*good information, i.e, technically sound information*".

Ao fazê-lo, essa pesquisa busca colaborar para a discussão sobre fatores e elementos das políticas de avaliação que contribuem para que as mesmas sejam úteis e, em conseqüência, para a formulação de novas avaliações ou realinhamento de políticas existentes.

O presente documento está organizado em quatro seções, além dessa Introdução e das Referências: o Marco Teórico, a Metodologia, os Resultados (Contribuições da Política de Avaliação da Aprendizagem) e as Considerações Finais. O Marco Teórico foi construído a partir dos conceitos de política pública e de avaliação, restrita à avaliação educacional. Em seguida, buscou os elementos que indicam qualidade nas experiências avaliativas e introduziu e discutiu o conceito de meta-avaliação para concentrar-se nos usos possíveis para as experiências avaliativas. Do modelo de meta-avaliação da avaliação de programas educacionais, a categoria Utilidade foi adaptada para análise de política de avaliação e a categoria Uso foi proposta. Na seqüência, o texto apresenta a metodologia empregada no estudo, iniciando-a pela contextualização e descrição da política foco da pesquisa em relato: a Avaliação da Aprendizagem (AA), no ciclo 2001 – 2004. A seção que apresenta os resultados é iniciada com o detalhamento da AA a partir dos elementos da categoria Utilidade, enriquecida por dados sobre a percepção das escolas quanto à utilidade esta política. Isso feito, são relacionados os usos relatados pelas escolas para os diversos elementos da política de avaliação. A última seção, Considerações Finais, apresenta uma reflexão sobre as contribuições (usos e utilidade) de políticas de avaliação em larga escala.

2. Marco Teórico

A construção do marco teórico para o presente trabalho resulta em duas categorias para análise de uma política de avaliação em larga escala, Utilidade e Usos. A primeira delas – Utilidade - está inserida no contexto dos padrões empregados para a meta-avaliação, conforme proposta do *Joint Committee on Standards for Educational Evaluation* (JCSEE), e a segunda foi elaborada a partir de uma discussão sobre usos possíveis da avaliação. O caminho para esta construção é descrito a seguir.

De início, faz-se uma breve apresentação das políticas públicas de avaliação e seu contexto para, em seguida, debruçar-se sobre o conceito de avaliação. O texto então focaliza a avaliação educacional[9] para, na seqüência, discutir seus padrões de qualidade. Isso posto, o marco teórico ora descrito concentra-se na conceituação e contextualização da meta-avaliação e de suas dimensões para, por fim, expor as principais tendências e discussões sobre a Utilidade e o Uso da avaliação, que dão o lastro para a construção do quadro de análise da presente pesquisa. As relações entre esses conceitos podem ser visualizadas no Mapa Conceitual apresentado no Apêndice 01.

2.1 As políticas públicas e o contexto da reforma do Estado

Nessa seção da fundamentação teórica, pretende-se situar a avaliação educacional como política no cenário mais amplo dos estudos das políticas públicas (*policies*), em especial aquelas com alguma proximidade com os movimentos de reforma de Estado e da Educação para Todos[10]. Ao fazê-lo, busca-se diferenciar avaliação de políticas, etapa do ciclo de análise das políticas públicas, de política de avaliação, em si um ciclo completo.

O conceito de política pública é polissêmico e engloba, a depender do teórico que o propõe, desde uma definição ampla, como “aquilo que o governo escolhe fazer ou deixar de fazer” (DYE: 1995: 02), até uma mais concreta como “programas de ação governamental visando a coordenar os meios à disposição do Estado e as atividades privadas, para a realização de objetivos socialmente

9 Para fins desse trabalho, entende-se avaliação educacional como avaliação da educação.
10 Quatro organismos internacionais vinculados à ONU (UNESCO, UNICEF, PNUD e Banco Mundial) patrocinaram, em 1990, uma conferência internacional em Jomtien, Tailândia, com o tema Educação para Todos. Os 155 governos presentes assinaram uma Declaração Mundial e um Marco de Ação comprometendo-se a assegurar uma

relevantes e politicamente determinados." (BUCCI, 2020:241 *apud* TEIXEIRA, 2006:20). Para fins desta pesquisa, entende-se, assim como Souza (2002: s/p), que programas e projetos do governo são considerados políticas públicas e sua implementação é "política pública posta em prática, transformação da política pública em ação".

Subárea das Ciências Políticas, o estudo das políticas públicas vinha sendo direcionado, em grande parte, aos processos e dinâmicas que levavam à sua formulação, ao passo que questões de implementação e de avaliação eram deixadas em segundo plano (FARIA, 2003). Entretanto, as políticas públicas, apesar de decididas pelos seus formuladores, são "traduzidas" principalmente ao nível da prestação do serviço, no contato da burocracia com o cidadão-usuário (LIPSKY, 1980). A etapa de implementação assume então um papel importante, não só pela operacionalização da política originalmente concebida, mas principalmente porque, quando e se alterada, uma nova política, derivada da anterior, é que será concretizada. No novo formato, nem sempre os objetivos traçados são atingidos ou considerados (RUA, s/d).

Para Walker (2004), a identificação de fatores que afetam a habilidade política de atingir seus objetivos tem sido uma preocupação presente na literatura sobre o tema. É possível distinguir uma série de abordagens de estudo sobre como se relacionam os formuladores e implementadores de uma determinada política. As teorias macro concentram-se no leque de respostas organizacionais às questões de implementação, enquanto os modelos que tratam do nível micro usam as interpretações e ações do ator local. Embora ambos os eixos teóricos estejam baseados em pressupostos diversos, eles oferecem visões muito mais complementares que excludentes sobre a política analisada (WALKER, 2004).

Nas duas últimas décadas, estudos – macro ou micro - sobre a implementação e avaliação das políticas vêm recebendo um olhar mais atento, em especial devido às mudanças mundiais ocorridas no cenário político. O aumento da demanda pelos serviços do Estado e o não acompanhamento correspondente na arrecadação de receitas provocaram a necessidade de enfoques mais racionais para os governos, semelhantes aos usados pelo setor privado (BOORSMA, 1997; OSBORNE, 1994). Além disso, foi possível observar o fortalecimento da sociedade civil que passa a exigir mais qualidade dos governos. Essas demandas e restrições formaram a base para uma crise que muitos governos responderam com uma proposta de reforma (OSBORNE, 1994), em um movimento batizado de Gerencialista.

---

educação básica de qualidade a crianças, jovens e adultos. O Brasil foi signatário desse documento e o reafirmou em Dakar, em 2000, em nova reunião da "Cúpula Mundial da Educação".

A lógica do Gerencialismo previa que um Estado com poucos recursos para atender a demandas sempre crescentes precisa ser "racionalizado". Para Osborne (1994), a solução para a crise não passa pela discussão de um governo maior ou menor: ela está centrada em um "melhor governo", empreendedor, flexível, com objetivo de maximização de produtividade e eficiência. Ou, como posto por Pimenta (1998), não é uma questão de reduzir, mas sim de melhorar a qualidade e eficiência do gasto público. Para tanto, o Estado precisa se tornar uma organização ágil.

Em grande parte dos países, a agilidade e a eficiência seriam obtidas por meio da modernização dos processos, destreza da burocracia e foco nos resultados. Em vários casos, isso significou privatização; em outros, a descentralização do governo federal pelas esferas estaduais e municipais ou ainda o fortalecimento das gestões locais.

Em que pese o reconhecimento da precariedade dos recursos dos Estados e seus Governos, nem todos os estudiosos concordam que a crise seja financeira. Os críticos do movimento encontram outras razões para as crises e, a partir delas, argumentam que as reformas são em verdade instrumentos a serviço de uma determinada linha político-ideológica, neoliberal (FLEURY, 1997). Ou, como discutido por Costa (2000:1), "a crise do Estado é, sobretudo uma crise de governabilidade, ou seja, a incapacidade de regular as relações entre economia e sociedade e os conflitos distributivos a elas inerentes, no contexto da ordem democrática e de um mundo globalizado". Já Santos (2001) levanta a questão da concentração de capitais e da sua "financeirização", o que tira do Estado a sua condição de principal investidor capitalista. Nessa linha de argumentação, há autores que associam os movimentos de reforma à crise do sistema capitalista que, sem espaço para expansão de seus mercados, invade o "território de atuação" do Estado, usando para isso o discurso neoliberal. Para reforçar essas linhas críticas, governos debatem-se, ainda hoje, com o agravamento dos mesmos problemas que os levaram a propor as reformas vinte anos antes. Entretanto, as ferozes críticas à proposta Gerencialista (ou ao "hegemônico modelo neoliberal") não impediram sua expansão.

O movimento de reforma de Estado atinge o Brasil principalmente com o governo de Fernando Henrique Cardoso (FHC). Bresser Pereira, um dos principais arquitetos do que considerou a mudança do Estado patrimonialista e burocratista para o Estado gerencial, defende "estratégias administrativas baseadas na ampla delegação de autoridade e na cobrança *a posteriori* dos resultados" (PEREIRA, 1998:9). A estrutura estatal precisou sofrer alterações para atender às proposições da reforma. No governo FHC, a administração pública deveria ser orientada para o

cliente e ter o foco direcionado para os resultados. Buscava-se a transparência e contava-se com informação; os procedimentos deveriam ser simplificados e as palavras chave eram autonomia e responsabilização (BRASIL, 2000).

O caminho da modernização, iniciado pela reforma fiscal, seria trilhado por um projeto de reforma complexo, através do qual se buscava, a um só tempo, fortalecer a administração pública direta e descentralizá-la através da implantação das agências executivas e organizações sociais controladas por contratos de gestão (BRASIL, 2000). Essas organizações sociais assumiriam serviços estatais que seriam publicizados (como saúde e educação), com isso simplificando-se fluxos e processos para a implementação das políticas. Ao governo caberia criá-las e regular sua implementação, ficando com a provisão direta de serviços como segurança territorial, por exemplo. Também caberia ao Estado a regulação na educação, na saúde, na cultura, no desenvolvimento tecnológico e nos investimentos em infra-estrutura, para compensar os desequilíbrios distributivos provocados pelo mercado globalizado e capacitar os agentes econômicos a competirem em nível mundial (PEREIRA, 1998).

Após os oito anos do Governo FHC, o Governo do Presidente Luis Inácio da Silva (Governo Lula) adotou a perspectiva de que o Estado, em lugar de ser a crise, deveria ser parte essencial da solução. Entretanto, manteve algumas das principais características do movimento Gerencialista, como o ajuste fiscal, a eficiência e a modernização da máquina estatal, ainda que pretendesse ir além do estado regulador. Quanto à máquina, propôs-se a manter o foco nos resultados para as organizações públicas. Para isso, precisaria também da sua otimização, com novos parâmetros de desempenho e incorporação de inovações tecnológicas; adequação dos quadros funcionais, reestruturação de carreiras e quadros de cargos com remuneração compatível com as responsabilidades e competências exigidas; o aprimoramento do atendimento ao cidadão (que deixa de ser o “cliente consumidor do serviço público”, do governo anterior, para ser considerado “membro de uma comunidade cívica, organizada e plena de direitos e deveres”), através da simplificação de processos, eliminação de exigências e controles desnecessários e facilitação do acesso aos serviços públicos; e a capacitação intensiva e permanente de servidores e dirigentes. E, assim como na administração anterior, o governo Lula previu o aprimoramento de sistemas de informações e o desenvolvimento de sistemas de monitoramento e avaliação de políticas, programas e projetos (BRASIL, 2003).

Em qualquer dos dois governos, os princípios de autonomia e descentralização que lastraram a ação administrativa implicaram a responsabilização dos gestores, no sentido da accountability11. Se antes a norma ditava o comportamento, impedindo a tomada de decisões ou a flexibilização da ação para o atendimento ao "cidadão-cliente", a mudança deveria se dar no sentido de maior fluidez de diretrizes e maior clareza aos propósitos, favorecendo o espaço para a ação. O foco saiu dos processos para se deter nos resultados. A regulação tenderia a deixar de ser fortemente burocrática para assumir o tipo "mercantil", segundo conceituação adotada por Afonso (2003), sem perda de poder para o Estado.

Nesse contexto de descentralização, otimização de recursos e foco nos resultados, uma maior atenção recaiu sobre os controles, auditorias e processos de *accountability* (BOORSMA, 1997; CASTANHAR; COSTA, 2002). Como visto nos discursos dos dois últimos governantes brasileiros, os sistemas de informação tornaram-se fundamentais para a tomada de decisão e para a prestação de contas. Daí o interesse por estudos de implementação e avaliação das políticas públicas. Daí também a necessidade de estratégias de avaliação de programas, embutidas em cada proposta de ação; da definição de sistemas de informação, como os censos escolares; e das políticas de avaliação com implementação em larga escala. É sobre a centralidade das políticas de avaliação, em especial de avaliação educacional, que trata a subseção a seguir.

2.1.1 A centralidade das políticas de avaliação

O discurso que se estabelece no panorama do Estado otimizado é calcado na racionalidade: a tomada de decisão é um processo racional, informado por dados levantados de forma rigorosa e sistemática. Além disso, o Estado precisa prestar contas aos cidadãos, por um lado, e garantir transparência, por outro. A avaliação é a ferramenta que permite que essas demandas sejam atendidas.

"Avaliar uma política é conhecer suas conseqüências" (DYE, 1995:320, tradução deste autor). Essas conseqüências podem ser analisadas sob diversas lentes: efetividade, eficiência, eficácia, impacto; com foco no processo ou no produto; podem ser estudadas individualmente, por programa, ou de maneira comparada, em estudos longitudinais ou transversais, contrastando um programa com outras experiências. Em qualquer das finalidades, as avaliações e pesquisas avaliativas devem ser realizadas de maneira sistemática e rigorosa, lançando mão de indicadores,

11 O termo *accountability*, de origem inglesa, tem sido muito utilizado em estudos de administração pública, relacionado, principalmente, às questões de prestação de contas à sociedade e de responsabilização do gestor público,

existentes ou criados por elas mesmas, e de critérios a partir dos quais os gestores públicos julguem os resultados encontrados (CASTANHAR; COSTA, 2002).

Contudo, para que sejam determinados indicadores e critérios em um processo avaliativo, há a necessidade de que, ainda na formulação, tenham sido estabelecidos metas e objetivos avaliáveis para as políticas públicas. Nem sempre essa definição ocorre. Muito frequentemente, as políticas são implementadas a partir de finalidades amplas, não traduzidas em metas (LIPSKY, 1980). O problema da indefinição das metas e objetivos para as políticas públicas tem impacto direto na determinação dos critérios e indicadores. Como obtê-los, de modo a que sejam capazes de informar quando um sistema ou uma unidade do sistema esteja ou não produzindo bons resultados (MOREIRA, 2002), se não há ainda o conceito do que seja aquilo que deve ser medido?

Essa tem sido talvez uma das razões para o paradoxo que se apresenta: ainda que as avaliações tenham obtido uma condição central no cenário político, seus resultados são pouco utilizados (SOUSA, 2003; HELENE, s/d). Além disso, os processos de formulação / tomada de decisões e de implementação das políticas são muito complexos e bem menos racionais que se pretende (Warde em entrevista a YAZBECK, 2007:18; SOUZA, 2002; WEISS, 1999; RUA, s/d). Esse quadro não é diferente no campo educacional que, no Brasil, vê a consolidação das avaliações e sistemas de avaliação educacional de abrangência nacional e, em paralelo, o crescimento das discussões sobre a falta de uso dos seus resultados.

No cenário de escassez descrito anteriormente, a área de Educação foi priorizada. Muitos governos, dentre os quais o brasileiro, entenderam que somente uma população educada teria condições de competir globalmente. Na América Latina, a partir da década de 1980 e mais intensamente nos anos 1990, foram implantadas reformas educacionais em diferentes países da região por meio de mecanismos considerados similares especialmente quando observadas as influências dos organismos supranacionais de fomento (MACHADO, 2007). De maneira geral, essas mudanças tiveram ênfase no ensino básico e visaram ao fortalecimento do "papel do Ministério da Educação como planificador e controlador da política educacional, bem como a delegação da operacionalização, execução e em grande parte da manutenção do ensino para as unidades subnacionais" (OLIVEIRA, 2002:70). As censuras foram e continuam sendo muito fortes à proposta de Estado regulador, no que é entendido como a transformação da educação em negócio (ANDRIOLI, 2002; GENTILLI, 1996; LIMA, 2002) e como a desresponsabilização do

---

inclusive legalmente. Para fins do presente trabalho, entende-se *accountability* como imputabilidade.

Estado de um serviço considerado essencial. Essas críticas tornam-se ainda mais presentes visto que, tantos anos depois, os indicadores de qualidade educacional continuam apontando para a não solução dos problemas e, em contraste ao discurso dominante, experiências asiáticas mostram melhorias educacionais sem que os países tivessem sofrido “reforma” (TEDESCO, 2003).

No Brasil, a base legal para a reforma na educação foi sendo estabelecida gradativamente, anterior ao governo FHC. O primeiro passo foi dado pela Constituição Federal de 1988 que apenas reconheceu a autonomia dos sistemas municipais de ensino. Oito anos depois, foi a Lei Federal 9.424/96 o que garantiu a implementação das políticas de descentralização, através da instituição do Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério (Fundef). Ainda em 1996, a Lei de Diretrizes e Bases 9.394/96 (LDB/96) redefiniria os papéis e responsabilidades de cada sistema de ensino, assegurando maior autonomia às escolas, flexibilização dos conteúdos curriculares e trazendo exigências de qualificação docente. Para complementar a base legal da reforma, em 1997, o governo federal lançou o Plano Nacional da Educação (PNE) com o objetivo de fazer cumprir as mudanças postas pela LDB/96. Vinte anos depois da promulgação da Constituição, a descentralização continuou dando o tom às políticas educacionais com a publicação, em 21 de junho de 2007, da Lei 11.494 (assinada um dia antes), que regulamenta o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação – Fundeb, uma expansão do conceito Fundef com a inclusão do Ensino Médio e da Educação Infantil.

Apesar de a base legal ter sido e continuar a ser favorável à descentralização e à autonomia dos municípios e das escolas, o que se percebe é que, na contramão desse discurso, foram adotadas várias políticas centralizadoras (CASTRO, 1999), como o lançamento das Diretrizes Curriculares Nacionais, em 1998, e das matrizes de referência do SAEB, divulgadas a cada nova aplicação de provas. Como posto por Oliveira (2002), o aparato de regulação e controle foi centralizado pelo Ministério da Educação (MEC), deixando pouco para as instâncias implementadoras (GENTILI, 1996). Além disso, sobre a descentralização ao nível municipal, os escândalos com o Fundef mostram o mecanismo não foi simples quando implementado. Acresce-se que o compromisso dos governos com os processos de descentralização não é confiável, ao longo dos seus mandatos, e há grande resistência da burocracia média aos tais processos, que afetam sua autoridade e poder (CORRALES, 2000). Especialmente ao nível da escola, a autonomia não foi concretizada, sobretudo quanto aos aspectos financeiros (que continuaram sob a mão dos gestores de sistemas) e, em muitos casos, também nas questões pedagógicas, já que nem sempre as escolas contavam e contam com um quadro docente capacitado. O discurso, no entanto, foi mantido e a cobrança por

melhores resultados também. Por essa razão, do mesmo modo que a avaliação teve um papel central no movimento geral de reforma do Estado, foi crucial para os movimentos educacionais, oferecendo um contraponto para o discurso de descentralização e de autonomia local e recebendo financiamentos generosos de governos e organismos financeiros (SANDER, 2002).

No Brasil de FHC, sob o comando do ministro Paulo Renato Souza, de 1995 a 2003, as políticas que levaram a avaliações nacionais se consolidaram em três níveis: o Fundamental, com o SAEB; o Médio, com o ENEM; e a Educação Superior, com o ENC (Provão), o Censo[12] e a ACE (CASTRO, 1999). Essas avaliações em larga escala não faziam parte de um projeto ou programa específico, constituindo-se elas mesmas políticas públicas a serviço da sociedade. Duas questões se colocam: que indicadores e critérios foram utilizados para que a avaliação produzisse informações para as políticas educacionais? E que indicadores e critérios devem ser utilizados para acompanhar as próprias políticas de avaliação?

Os indicadores e critérios utilizados pela avaliação para informar sobre a qualidade da educação no Brasil têm sofrido, como no ambiente maior da política pública, a dificuldade de delimitação, visto que, como já discutido, os objetivos das políticas educacionais nem sempre são ou foram traduzidos em metas. O que se pretende, hoje como no passado, em termos de reforma educacional, é uma educação de qualidade para todos. A definição do que seja uma educação de qualidade está longe de ser feita. Por exemplo, a LDB 9.394/96 pretende a formação de cidadãos que possam ser inseridos no mercado de trabalho, o respeito à diversidade cultural e a contribuição para o contexto social. Como medir o atingimento do que está posto na Lei, dentro das limitações técnicas e orçamentárias e das constrições de tempo que caracterizam os governos? Como definir indicadores que abarquem desde elementos de construção de cidadania, ao longo do período de escolarização formal, e ao mesmo tempo fazê-lo de tal maneira que possam ser levantados? Que indicadores poderiam ser usados como *proxy*? Os dados a serem levantados seriam apenas aqueles que falam do atingimento dos objetivos ou deveriam abranger possíveis razões para desvios encontrados?

Use-se o exemplo do SAEB, cuja primeira aplicação foi feita em 1990 e que sofreu uma grande reformulação metodológica em 1995, para permitir comparações ao longo do tempo, a partir de quando se consolida. Entretanto, suas matrizes de referência (que chegaram a ser chamadas, inicialmente, de matrizes curriculares) trazem indicadores que tratam da aquisição das

12 Levantamentos de dados e estatísticas não são, isoladamente, avaliação. O Censo, nesse contexto da Avaliação da Educação Superior, é um elemento da política quando articulado com as demais funções do sistema avaliativo.

competências e habilidades pertinentes a um determinado nível de ensino, sem sequer tangenciar as questões de cidadania e diversidade. Ainda assim, esse é o sistema que monitora a qualidade da Educação Básica no Brasil. Como, então, analisar os resultados do SAEB? Sob o olhar da educação nacional ou sob o foco das aplicações de provas de Língua Portuguesa e Matemática?

Por outro lado, pesquisadores como Sousa (2003) argumentam que o formato das avaliações impacta o currículo e a cultura organizacional da instituição escolar, reduzindo-os ao que é endereçado pelos instrumentos da avaliação. Nesse sentido, os efeitos acabam sendo negativos em termos de qualidade educacional. As abordagens avaliativas adotadas, não só pelo SAEB, têm sofrido críticas constantes, por serem consideradas muito limitadas, ou por se tornarem ferramenta de controle da educação como formadora de mão de obra para o mercado (GENTILLI, 1996), ou ainda por terem um caráter quantitativista, muito restrito para informar sobre a qualidade educacional pretendida. Os delineamentos em larga escala sempre apresentarão um recorte da realidade, por mais que tentem abordá-la de maneira compreensiva, de resto como qualquer pesquisa. Quais são, então, os indicadores e critérios utilizados para avaliar o próprio SAEB? O Sistema não traz uma indicação de meta-avaliação, apesar de, obviamente, ser objeto de estudo de vários grupos de pesquisa no Brasil.

As críticas feitas não impediram o SAEB de se consolidar e de impactar as políticas ao nível estadual e municipal[13]. Como posto por Schwartzman e Verhine, pior do que pouca informação é ter informação nenhuma, o que justifica as avaliações cujos delineamentos tenham sido resolvidos de maneira apropriada (SCHWARTZMAN, 2005; VERHINE, 2008). O sucesso do SAEB motivou os Estados a lançarem suas próprias experiências de avaliação ou a consolidarem as existentes, especialmente a partir de 2000. Alguns, como o Ceará, o fizeram por meio da aplicação censitária de provas formuladas com itens do SAEB nos anos pares (já que a prova do SAEB era aplicada em anos ímpares); outros, como São Paulo e Bahia, criaram sistemas próprios de avaliação. A partir de 2005, com a criação do ANRESC (Avaliação Nacional do Rendimento no Ensino Escolar, conhecida como Prova Brasil), muitas dessas experiências desapareceram ou se transformaram, com perda de massa crítica nos Estados e uma volta à concentração das avaliações pelo INEP.

O ANRESC, assim como o SAEB, sofreu com a falta de metas das políticas públicas educacionais brasileiras. Essa ausência foi, em parte, sanada por meio do Decreto nº. 6.094 (de 24 de abril de

13 Para um panorama sobre os programas de avaliação conduzidos no Brasil ao nível estadual, ver BONAMINO; BESSA; FRANCO (Org), 2004.

2007), que regulamenta o Plano de Metas Compromisso Todos pela Educação, proposto pela sociedade civil e adotado pelo Governo Federal. Um grande avanço, nesse Decreto, é a definição do IDEB – Índice de Desenvolvimento da Educação Básica –, criado pelo INEP a partir de dois outros indicadores: um de fluxo e outro de desempenho, como indicador de acompanhamento do plano. Mesmo assim, há claros *gaps*: por exemplo, uma das metas é ter todas as crianças alfabetizadas até os oito anos. Como o indicador de desempenho vem da Prova Brasil, aplicada a alunos da 4ª série e, portanto, mais velhos, o IDEB como posto hoje não dará conta de informar sobre a alfabetização. Já a Provinha Brasil, aplicada aos alunos de alfabetização, não compõe o IDEB, embora isso possa ser facilmente corrigível, com a criação de um IDEBinho. O Decreto nº. 6.094, no qual a avaliação é componente, está no contexto do PDE _ Plano de Desenvolvimento da Educação, lançado pelo ministro Fernando Haddad em 2007, como ferramenta de operacionalização do novo Plano Nacional da Educação (PNE). O PDE reúne mais de 40 programas e projetos a serem desenvolvidos em regime de colaboração com estados e municípios. Também para o Plano o IDEB é o indicador, com a meta de um IDEB Brasil igual ou superior a 6,0 no ano de 2021.

O Decreto nº. 6.094 é interessante não só pela determinação do indicador, mas também pela definição da articulação entre governo federal e demais esferas, no sentido de apoiá-las no caminho para o cumprimento das metas. Durante muito tempo, os responsáveis pelo SAEB (LOCATELLI, 2001), bem como gestores de outras avaliações em larga escala, têm argumentado que há necessidade de articulação das políticas de avaliação com outras políticas que favoreçam a mudança. O que se tem percebido, ao longo do tempo, é que há uma expectativa de que a mudança ocorra como conseqüência direta da divulgação dos resultados da avaliação, sem que intervenções sejam realizadas com essa finalidade. As políticas de avaliação são isoladas, como se trouxessem um fim em si mesmas.

Definir o(s) indicador(es), por mais restrito(s) que seja(m), é um passo importante para o acompanhamento das políticas educacionais já que o quadro educacional no Brasil continua extremamente problemático após tantos anos de uma reforma que considera a educação uma área prioritária. Desde os meados dos anos 90, o mais grave problema, no Brasil, não é mais o atendimento à população de 7 a 14 anos (96,5% em 1998 e 97,0% em 2005) ou mesmo a taxa de analfabetismo para a faixa 15 anos ou mais (13,5% em 2000), segundo dados do INEP. A exclusão e a discriminação passam a ser associadas à baixa qualidade da educação ofertada no País, que tem sido atribuída, muitas vezes, à expansão do sistema educacional (CASTRO, 1999), em um cruel mecanismo de troca da exclusão (por restrição da oferta de vagas) por mais exclusão

(pela baixa qualidade). A baixa qualidade é traduzida por altas taxas de repetência, de abandono e evasão, de analfabetismo funcional, dentre outras. Indicadores de desempenho em Língua Portuguesa e em Matemática na Educação Básica e, mais recentemente, o IDEB mostram que o aluno brasileiro tem perdido tempo na escola em lugar de aprender as competências e habilidades mínimas para prepará-lo para a vida profissional. A publicação dos dados do IDEB 2007[14], quando comparados com os dados do IDEB 2005, mostra um movimento positivo muito débil, tanto para o Ensino Fundamental (3,5 para 3,8) quanto para o Médio (3,4 para 3,5), especialmente se considerada a escala de 0 a 10. O fato é que, em 2008, 17 anos após a primeira aplicação do SAEB, os resultados são tão preocupantes quanto em 1990.

Diante do quadro apresentado, entendendo programas e projetos como política pública em ação e categorizando as avaliações em larga escala como políticas públicas centrais em um Estado Regulador, é importante que estudos e pesquisas sejam conduzidos e que incluam, além da formulação, as etapas de implementação e de avaliação de um ciclo. Em relação a políticas de avaliação educacional, não há muitos relatos sobre sua implementação ou avaliação no Brasil. Na última década, apenas o governo do Rio de Janeiro havia contratado uma meta-avaliação para o sistema de avaliação implementado. Mais recentemente (2005), tem havido uma série de estudos, em parte contratados pelo INEP, sobre suas avaliações. Esses estudos, no entanto, são fragmentados: alguns pesquisadores tratam das análises, enquanto outros discutem instrumentos, e assim por diante. Seria interessante um esforço de uma meta-avaliação global das políticas de avaliação. Seus achados poderiam ajudar a compreender essas políticas e talvez a potencializar sua utilização na busca pela reversão do quadro educacional brasileiro.

Na presente pesquisa, optou-se por uma análise das contribuições (utilidade e usos) de uma política de avaliação implementada na Bahia entre 2001 e 2004. Na composição do quadro de análise dessa política, é importante que os conceitos de avaliação e de meta-avaliação sejam definidos. É disso que tratam as subseções a seguir.

14 Dados disponíveis no http://ideb.inep.gov.br/*Site*/. Acesso em 14.07.2008.

2.2 O conceito de avaliação educacional e os padrões que indicam sua qualidade.

O conceito de avaliação, quando atrelado à educação, existe desde que a instituição escola foi criada, mas foi na década de 1940 que começa sua formatação para o que se conhece nos dias de hoje (VIANNA, 2000). Desde então tem sido modificado, respondendo, por um lado, aos avanços metodológicos e tecnológicos que permitem, por exemplo, aumento simultâneo do escopo sob avaliação, e por outro, às mudanças dos paradigmas científicos que implicam alteração nos interesses pelos objetos e, conseqüentemente, no foco avaliativo.

Para a construção do quadro de análise para o presente trabalho, foi fundamental definir avaliação para que fosse possível discutir meta-avaliação e, a partir dela, utilidade e uso da avaliação. Com esse objetivo em mente, foi preciso restringir a avaliação ao campo da educação e, para conceituá-la, relacionar algumas de suas aplicações, que a adjetivam; fazer um paralelo da avaliação com a pesquisa científica, diferenciando-as; para então discutir seus padrões de qualidade. Para tanto, as próximas subseções: 1) conceituam avaliação educacional de modo geral; 2) adjetivam a avaliação, inclusive distinguindo a avaliação na escola da avaliação da escola (em um recorte de política pública, com aplicação em larga escala) para 3) comparar avaliação e pesquisa e, na seqüência, 4) discutir alguns aspectos a considerar quando do delineamento de uma avaliação em larga escala em 5) respeito aos padrões de qualidade da avaliação.

2.2.1 O conceito de avaliação educacional

O conceito de avaliação esteve inicialmente restrito à mensuração do desempenho (ou do rendimento escolar) e à verificação do cumprimento dos objetivos curriculares, como posto por Tyler nos anos 30-40. Em 1967 Scriven já associava a avaliação ao julgamento de valor de um objeto "para uma certa destinação" (SCRIVEN *apud* VIANNA, 2000:25)[15]. A idéia de uma sistematização da coleta de dados para que esse julgamento de valor fosse feito influenciou uma série de avaliadores, dentre os quais Stufflebeam (1974), autor base para a discussão de meta-avaliação, e Luckesi (2000), de quem o presente trabalho adota parcialmente o conceito de avaliação. Nessa linha de argumentação, o julgamento de valor está atrelado a uma posterior tomada de decisões, como visto em Sousa, que conceitua a avaliação como "um processo de

15 Para um panorama sobre a evolução do conceito da avaliação, consultar DIAS SOBRINHO, 2003:13-52, WORTHEN; SANDERS; FITZPATRICK ,2005:33-59, e VIANNA, 2000, em seus seis primeiros capítulos, nos quais discorre sobre Tyler, Cronbach, Scriven,Stufflebeam e Stake, avaliadores e pesquisadores que contribuíram enormemente para a avaliação como é entendida na atualidade.

busca de compreensão da realidade estudada, com o fim de subsidiar a tomada de decisões quanto ao direcionamento das intervenções" (SOUSA 1987 *apud* ABRAMOWICZ, 1994:95).

Mais recentemente, o JCSEE (1994) definiu avaliação educacional como uma análise sistemática do valor ou do mérito de um objeto educacional[16], inserindo na conceituação as necessidades dos *stakeholders*, diretamente relacionadas ao valor. Não há, portanto, um conceito moldado e pronto do que seja avaliação, mas, no geral, percebe-se um consenso sobre a coleta de dados de modo a propiciar que um julgamento seja feito, preferencialmente visando à tomada de decisões que resulte no melhoramento do objeto sob avaliação.

Nesse trabalho, avaliação é compreendida, em acordo com o proposto por Luckesi (2000) em discussão sobre avaliação da aprendizagem escolar, como "um juízo de qualidade sobre dados relevantes, tendo em vista uma tomada de decisão". Utilizam-se as palavras daquele autor para esclarecer o conceito:

> Em lógica, juízos são afirmações ou negações sobre alguma coisa. Essas afirmações ou negações poderão incidir sobre o aspecto substantivo ou sobre o aspecto adjetivo da realidade. O juízo que se faz sobre o aspecto substantivo da realidade recebe a denominação de *juízo de existência*, na medida em que a sua expressão pode ser justificada pelos dados empíricos da realidade. O juízo, porém, que expressa a qualidade do objeto que está sendo ajuizado recebe a denominação de *juízo de qualidade*, desde que incida sobre uma realidade atribuída ao objeto. O primeiro pretende dizer o que o objeto é; o segundo tem por objetivo expressar uma qualidade que se atribui ao objeto. Enquanto o juízo de existência é produzido numa relação direta do sujeito com o objeto, o juízo de qualidade é produzido por um processo comparativo entre o objeto que está sendo ajuizado e um determinado padrão ideal de julgamento (LUCKESI, 2000:69).

O padrão de julgamento – seja na sala de aula, seja em larga escala – pode ser estabelecido em uma comparação com um padrão entendido como ideal e comum e, nesse caso, tem-se uma avaliação referenciada a critério. Quando o padrão é estabelecido na comparação entre os resultados daqueles sob avaliação, diz-se da avaliação referenciada a norma.

Ainda segundo Luckesi sobre a avaliação da aprendizagem:

> A segunda variável a ser considerada na avaliação é que o juízo de qualidade deve estar fundado sobre *dados relevantes da realidade*. A qualidade de um objeto não lhe será atribuída ao bel-prazer de quem o julga, mas sim a partir de caracteres que este determinado objeto possua. [...] É um juízo de qualidade, porém não uma qualidade arbitrária, mas sim uma qualidade que está fundada em propriedades "físicas" dessa mesma realidade. Propriedade "física", aqui, está sendo entendida como caráter efetivo e objetivo da realidade a partir do qual se pode estabelecer a qualidade desse objeto. No

16 "*Evaluation is the systematic assessment of the worth* **or** *merit of an object*" (JCSEE, 1994:03, grifo desse autor). Essa ampliação conceitual é interessante porque se percebe, ao contrário de uma discussão grande da atualidade , que uma avaliação pode apenas interessar-se pelo mérito de um objeto, sem dedicar-se a levantar seu valor (ainda que não seja possível levantar o valor sem ter antes assegurado o mérito). A definição do JCSEE utiliza a conjunção *or* em lugar de *and*.

> caso da aprendizagem, as propriedades "físicas" são as condutas aprendidas e manifestadas pelos alunos (LUCKESI, 2000:7-71).

Utilizando a definição de Luckesi como base para a conceituação da avaliação educacional em larga escala, relacionam-se essas propriedades do objeto, usualmente, à matriz de referência da avaliação. Em algumas experiências de avaliação educacional, como o próprio SAEB no início, essa matriz é chamada de matriz curricular ou ainda de referência curricular.

O terceiro elemento da conceituação-base nas palavras do pesquisador:

> O terceiro elemento que compõe a definição de avaliação é a *tomada de decisão*. Um juízo de existência encerra-se na afirmação ou na negação do que um determinado objeto é; no caso do juízo de qualidade, ao contrário, implica alguma coisa a mais, implica uma tomada de posição, um estar a favor ou contra aquilo que foi julgado. Sendo o juízo satisfatório ou insatisfatório, temos sempre três possibilidades de decisão: continuar na situação em que se está, introduzir modificações para este objeto ou situação se modifique para melhor; ou suprimir a situação ou o objeto (LUCKESI,2000:71).

Avaliação educacional é, para fins do presente trabalho, a busca de objetivação do julgamento sobre uma determinada realidade, capturada a partir de um recorte daquilo que é entendido como "dados relevantes". Há, portanto, duas decisões cruciais que fundamentam um processo avaliativo: a determinação do que seja "dado relevante" e a definição de um mínimo necessário que defina a qualidade do objeto e que, nesse sentido, permita que a mesma seja julgada. Em outras palavras, é essencial a determinação do que seja minimamente aceitável para diferencia-lo daquilo que não atinge tal patamar. Essa é exatamente a característica da avaliação que a torna uma ferramenta política, tanto mais impactante quanto maior for sua abrangência. A figura a seguir mostra a operacionalização do conceito aqui adotado.

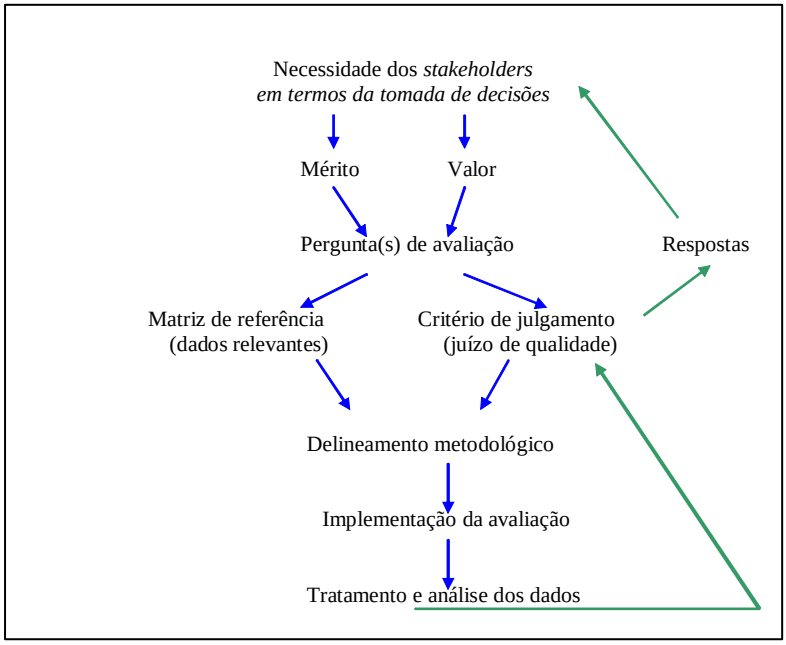


Ilustração 1: Conceito de avaliação atrelado à definição de dados relevantes e ao julgamento de qualidade.

Na literatura disponível, é interessante perceber que a discussão sobre avaliação tem sido muito voltada para a incapacidade de seus instrumentos captarem a realidade quando, em verdade, os argumentos seriam mais enriquecedores se, em lugar de tanta ênfase nos instrumentos, uma atenção maior fosse dada às matrizes de referência (ou, em muitos casos, à ausência das mesmas) e aos critérios utilizados para o julgamento. Há também um esforço enorme na adjetivação da avaliação, conforme sua finalidade, aplicação temporal ou abordagem, como pode ser visto na subseção a seguir.

2.2.2 A adjetivação da avaliação

A avaliação tem recebido, ao longo do tempo, várias adjetivações. Para exemplificação[17], apresentam-se seis diferentes categorizações, em termos de: 1) tempo de realização, 2) finalidade, 3) objeto da avaliação, 4) participação dos avaliados e outros *stakeholders* na formulação e implementação, 5) impacto para os avaliados para, por fim, 6) fazer-se a distinção entre avaliação na escola da avaliação da escola, fundamental para o recorte da presente pesquisa. Algumas vezes, as adjetivações se repetem nas diversas categorias, como será visto nos próximos parágrafos.

Na categorização da avaliação pelo tempo de ocorrência, é possível identificar:

a) a avaliação diagnóstica, realizada durante o planejamento ou no início de uma determinada intervenção, com fins de informar sobre o *status* do objeto avaliado no tempo zero e de, ao fazê-lo, favorecer a aproximação do planejamento com a realidade a ser impactada, dessa maneira colaborando para seu refinamento;

b) a avaliação de processo, concretizada durante a implementação da ação, cujo objetivo é contribuir para o atingimento das metas por informar os decisores sobre o que está dando certo e o que não está; e, por fim,

c) a avaliação de produto (ou de desempenho), normalmente realizada ao final do período de intervenção, com o objetivo de verificar o cumprimento das metas. Relacionada ao tempo, também é possível encontrar a distinção entre uma avaliação transversal, cujo objeto é observado em um momento único e definido, da avaliação longitudinal, cujo objeto é acompanhado ao longo de certo período, em coletas de dados subseqüentes que favoreçam a análise, de maneira comparável, das mudanças ocorridas.

Mais voltada para a finalidade da avaliação, a categorização em avaliação somativa e em avaliação formativa foi proposta em 1967 por Scriven. A primeira ocorre ao final do processo na

17 Para um aprofundamento nas categorizações da avaliação, consultar DIAS SOBRINHO, 2003: 29-52.

investigação do cumprimento dos objetivos para os quais a ação foi delineada e do seu mérito. Essa modalidade serve aos decisores, por exemplo, na resolução sobre a continuidade da ação, da sua modificação ou ainda da sua não repetição em versões futuras. Já a avaliação formativa visa basicamente o aperfeiçoamento do objeto avaliado durante sua implementação. Essa discussão, quando trazida para a sala de aula, traz uma outra categorização: avaliação da aprendizagem (no processo) e a avaliação de desempenho (ou de rendimento), realizada ao final do curso. É o próprio Scriven (1967 *apud* LETICHEVSKY *et alii*, 2007:449) quem defende o caráter complementar dessas duas modalidades avaliativas.

Quando observado o objeto da avaliação, pode-se diferenciar, por exemplo, a avaliação de produto, voltada para o produto de uma determinada intervenção, da avaliação de impacto (que visa observar as diferenças no objeto avaliado entre o tempo zero da intervenção e o tempo final) e da avaliação de efetividade, relativa à investigação do cumprimento dos objetivos maiores, sociais, da ação avaliada.

Quando o elemento de caracterização é o modo de elaboração e implementação da avaliação, há mais recentemente uma série de autores que defendem a avaliação participativa, na qual os *stakeholders* têm voz desde o planejamento da ação até sua avaliação final, e mesmo a avaliação emancipatória, que visa dar maior poder ao avaliado ao convidá-lo a repensar a realidade e a tomar decisões sobre sua própria ação e posterior avaliação (SAUL, 2001). Ao se envolver o *stakeholder* (principalmente os avaliados) no processo de planejamento e implementação da avaliação, espera-se uma maior utilização dos seus resultados, especialmente ao comparar essa abordagem com a estratégia mais tradicional. Nesta, a equipe de avaliadores assume um papel ativo, enquanto os *stakeholders* são colocados em passividade (PATTON, 1997). Além disso, a avaliação pode exercer um papel (como ocorre com a Matriz do Marco Lógico[18]) de ajudar a coerência e a coesão do grupo, por apontar os aspectos importantes, ou, como posto por Greene (1988 *apud* WEISS, 1998:25), impactar a aprendizagem do grupo e seu posicionamento frente à intervenção, a ponto de revigorar sua prática. Essa discussão sobre avaliação participativa é muito interessante quando se pensam nas experiências em larga escala, em caráter nacional ou mesmo local, envolvendo, às vezes, milhares de pessoas. Nesse cenário, a preocupação é exatamente

18 Para uma leitura breve sobre marco lógico, consultar BROSE, 2001:279-286 e RUA, s/d. A aplicação prática do marco lógico. Disponível em www.enap.gov.br/downloads/ec43ea4fAvaliacao_pratica_marco_logico.pdf. Acesso em 18.03.2008. De acordo com Brose, o marco lógico é um instrumento de gestão de programas e projetos públicos, com potencial de aperfeiçoamento dessa gestão e, desde os anos 80, tem tido utilização freqüente no Brasil (BROSE, 2001).

definir o que significa *participativo*, quem participará e de que forma o indivíduo, o grupo ou a organização serão envolvidos.

Em relação ao impacto da avaliação sobre os avaliados, uma adjetivação da avaliação é preciosa especialmente quanto aos controles: avaliação *high stakes* e avaliação *low stakes*. Diz-se de uma avaliação *high stakes* quando seu processo e resultados afetam diretamente o indivíduo, a organização, o programa. É, por exemplo, o caso do vestibular como meio para ingresso ao nível superior. A avaliação é *low stakes* quando seus efeitos não têm tal impacto para os avaliados, como ocorre no SAEB. Quanto mais *high stakes* for uma avaliação, tanto maior tende a ser a necessidade de controle sobre as informações coletadas.

Antes de concluir essa subseção, uma categorização da avaliação é importante para o recorte do presente trabalho: a avaliação do sistema educacional (por vezes chamada avaliação da escola) difere daquela avaliação encampada pela própria escola (avaliação na escola).

A primeira – avaliação do sistema educacional - tende a ser administrada em larga escala e, no Brasil, tem sido concretizada como política pública, normalmente voltada para a regulação dentro do contexto de Estado já discutido, tendo como seus decisores os gestores dos sistemas educacionais. Seus dados relevantes são postos em matrizes de referência que, por vezes, focalizam competências curriculares e, em outras ocasiões, incluem outros indicadores de sistema, como taxas de fluxo e de eficiência. Esse tipo de avaliação pode ser entendido como uma série de procedimentos, previamente estabelecidos, para coleta e tratamento de dados (relevantes) coletados em larga escala que, analisados a partir de padrões e critérios relacionados aos objetivos do sistema educacional, espera-se sejam usados para informar a tomada de decisões e para favorecer um julgamento de valor. Muitas vezes implementada por agentes externos ao sistema educacional, essa avaliação é também conhecida como avaliação em larga escala.

Ressalte-se que se a avaliação concentrar-se-á nas perguntas sobre o mérito ou se estará voltada também para o valor do objeto dependerá da "encomenda" avaliativa. Nos dois casos, ela é uma ferramenta poderosa em um contexto no qual, de maneira geral, os governos buscam eficácia[19],

19 Eficácia – refere-se ao resultado do sistema, em geral associado ao desempenho dos alunos, ao final de cada ciclo, quanto a competências e habilidades que deveriam dominar. Eficiência – diz respeito à otimização dos recursos durante a implementação das ações, de modo a permitir o atingimento das metas e objetivos. Eqüidade – relacionada à capacidade do sistema de garantir o melhor desempenho possível para seus alunos, independente de sua origem sócio-econômica.

eficiência e eqüidade para seus sistemas e precisam de informações que lhes ajudem a definir ou a alterar políticas educacionais que levem à melhoria da qualidade da educação que oferecem.

A avaliação na escola, diferentemente, é conduzida por professores e coordenadores educacionais e visa a atender pais, professores, diretores, além dos próprios alunos. Como restringe-se à unidade escolar, essa avaliação é realizada em pequena escala, especialmente quando comparada àquela avaliação do sistema educacional, que tende a abranger o conjunto de escolas e outros elementos.

Para a avaliação na escola, dados relevantes estão associados aos conteúdos conceituais, procedimentais e atitudinais (em tipologia oferecida por Zabala, 1998) que fazem parte do currículo das disciplinas e do plano de curso. As decisões a serem tomadas estão voltadas para o processo ensino x aprendizagem. A avaliação é conduzida em multiplicidade de instrumentos e grande freqüência. Para essa conceituação, entende-se que a tomada de decisão serve a um compromisso ético com a aprendizagem do aluno, como posto por Demo (2002), com a aprendizagem do próprio professor, e que, também, serve a propósitos de certificação cujos resultados, em termos institucionais, são traduzidos em aprovação e reprovação dos alunos na disciplina ou curso.

Reconhecendo a importância da avaliação na escola e na sala de aula, este trabalho, no entanto, preocupou-se com as políticas de avaliação que resultam em avaliações em larga escala e focalizou as contribuições da política de Avaliação da Aprendizagem, implementada na Bahia entre 2001 e 2004. Para descrever a política-foco (ver Subseção 3.1), foram utilizados alguns dos adjetivos aqui discutidos.

Cada uma das avaliações, nas suas diferentes adjetivações, pressupõe um delineamento específico, ainda que todas devam ser idealizadas em paralelo ao planejamento da própria ação a ser avaliada. Para fins deste estudo, o que define a avaliação é a busca por dados relevantes que permitam um julgamento de valor sobre o objeto avaliado. Para tanto, ela necessariamente passa por levantamento sistemático e rigoroso de informações, à moda da pesquisa científica, de quê se aproxima, mas com o quê não deve ser confundida, como pode ser visto na subseção a seguir.

2.2.3 Avaliação não é pesquisa científica

A necessidade de rigor na coleta de dados, a demanda por um modelo teórico que a guie, a busca por um recorte de realidade que possibilite a compreensão e o julgamento de um determinado objeto, a procura por uma abordagem metodológica que viabilize a resposta às perguntas avaliativas aproximam a avaliação da pesquisa científica e fazem com que a primeira, ao longo do tempo, tenha se aproveitado do aprimoramento da segunda. As melhorias dos instrumentos e técnicas de coleta de dados e o aperfeiçoamento dos métodos de análise de dados coletados, possibilitando, inclusive, seu estudo simultâneo em diversos níveis (como a sala de aula, a escola e o sistema), levam muitos pesquisadores a confundir avaliação com pesquisa científica.

Há, no entanto, diferenças cruciais entre elas. Como posto por Cronbach (1977 *apud* STAKE, 1982:14), a definição da avaliação como uma atividade científica nos leva a ignorar aspectos significativos da área, além da adoção de critérios falsos de excelência. Ainda segundo Cronbach, a avaliação é uma atividade política, uma função no sistema social. A definição da avaliação como atividade política (e não científica) é demonstrada nesse trecho de Cooksy e Caracelli, ao discutir preocupações da meta-avaliação:

> Meta-avaliação é avaliação, embora tenha a avaliação como seu objeto (*evaluand*). Dessa maneira, está sujeita a todas as mesmas considerações políticas de qualquer avaliação. Pressões políticas e comerciais nos clientes da avaliação ou sobre seus avaliadores podem resultar em avaliações indefensáveis ou em acusações de baixa qualidade de processo avaliativo (Chelimsky, 1987; Leeuw, 2003; Schwartz, 1998; Weiss, 1973; Wildavsky, 1972). A adequação técnica da avaliação e os esforços para o atendimento aos usuários da avaliação são igualmente importantes para as políticas e valores competitivos nos quais as avaliações estão caracteristicamente inseridas (Greene, 1990; Patton, 2003). Se uma síntese avaliativa leva a resultados impopulares sobre uma determinada intervenção, a adequação técnica da meta-avaliação provavelmente será questionada. Se uma meta-avaliação conduzida com objetivo de investigar a capacidade avaliativa de certa organização tem como achado que algumas de suas unidades demonstram capacidade inferior às demais, uma percepção de vencedores e perdedores pode ser criada, o que minaria os esforços para melhorar a capacidade avaliativa como um todo. Assim como em qualquer avaliação, não há uma resposta simples a essas preocupações na meta-avaliação. (COOKSY; CARACELLI, 2005:35-36, tradução deste autor)

O caráter político da avaliação implica um árduo processo de negociação e, nesse sentido, os delineamentos precisam de flexibilidade de modo a que se possa atender à demanda daqueles que a "encomendam" (REBOLLOSO *et alii*, 2002; DIAS SOBRINHO, 2003). Por essa razão, não há uma verdade única a ser imposta aos demais, mas sim uma constante procura de reconciliação entre os aspectos de qualidade técnica e necessidades dos *stakeholders*. Em pesquisa científica,

também não há uma verdade única, mas a busca se volta para o rigor metodológico que resulte em conhecimento. Como posto por Weiss (1998) e por Ferrer (1997), a interação com aqueles interessados na avaliação ocorre desde o início do planejamento, quando se busca corresponder aos anseios dos *stakeholders* com um delineamento avaliativo que os atenda, e ao final do processo, ajude-os a compreender (e a utilizar) os resultados. Ainda que essa interação possa ocorrer em pesquisa científica, não a define.

Além disso, há uma diferenciação na avaliação entre cliente e demais *stakeholders*, que são os usuários e os interessados na avaliação, que, normalmente, afasta a avaliação da pesquisa. Aquele que encomenda o processo avaliativo (o cliente) é, por natureza, um *stakeholder* de grande importância e garante um bom tempo de negociação com o avaliador para o atendimento de suas necessidades. Em pesquisa científica, raramente há essa demanda por negociação com o cliente; é mais freqüente a adequação da pesquisa às linhas de fomento nos diversos órgãos de financiamento. A subseção que trata da categoria Utilidade aprofundará a conceituação de cliente (ou contratante), *stakeholders* e usuários.

As diferenças entre avaliação e pesquisa científica não se resumem apenas ao contexto político que está na essência da primeira, mas são percebidas principalmente no que concernem as suas finalidades. Como argumentado por Abramowicz, a avaliação não está preocupada com a produção de conhecimento ou de generalizações, mas com a "utilidade imediata do conhecimento produzido" (ABRAMOWICZ, 1994:91). Essa posição é também defendida por Weiss, ao propor que o objetivo maior da avaliação é contribuir para que pessoas e organizações aprimorem seus planos, políticas e práticas visando ao bem-estar geral. Weiss acredita que, dentre os avaliadores, haja alguma motivação para a contribuição para o conhecimento geral, mas que suas expectativas estejam voltadas para afetar o modo com que as agências governamentais e instituições, de modo geral, endereçam os problemas da sociedade (WEISS, 1999).

Na década de 90, autores como Cook, Patton, Shadish, Cook ou Leviton (*apud* REBOLLOSO *et alii*, 2002:15) consideraram que havia uma crise de utilização da avaliação. Esses autores reforçaram a tese de que, para além da perspectiva tradicional da qualidade (da pesquisa) baseada na objetividade e na validade, a avaliação tinha um impacto claro em termos da solução de problemas e melhoria de condições sociais e organizacionais e deveria, portanto, associar sua qualidade à utilidade para aqueles com interesse no processo avaliativo. Esse caráter utilitário da avaliação contribui para diferenciá-la da pesquisa científica. Em última instância, as pesquisas científicas tendem a buscar o bem estar geral e as pesquisas aplicadas investigam possibilidades

de utilização mais imediata do conhecimento. Nesse sentido, estão próximas da avaliação. Sua relação com a base teórica e com a contribuição para a construção do conhecimento, como característica e objetivo maior, entretanto, é propensa a distanciá-las. A lonjura tende a permanecer quando a avaliação é implementada como uma política.

Se avaliação não é pesquisa científica, como então delineá-la? Visto que o foco do presente trabalho é uma política de avaliação, a próxima subseção discute aspectos de delineamento da avaliação como política para, em seguida, dedicar-se aos padrões e aos critérios que assegurem sua qualidade.

2.2.4 O delineamento de uma política de avaliação

O desenho de uma política de avaliação, para implementação em larga escala, precisa considerar três etapas, cada uma interferindo nos resultados da etapa anterior, até que um projeto consistente possa ser visualizado[20]. A discussão dessas etapas, em um caráter prático (mais que teórico), favorece a transposição do conceito da avaliação para o contexto de política pública. Assim como acontece no delineamento de uma pesquisa científica, a delimitação do objeto, a definição da metodologia e as implicações do processo e de seus resultados devem ser ponderadas e amadurecidas, ainda que, ao final, não se possa dizer de um delineamento pronto e acabado. A implementação, se monitorada, permitirá um retorno ao desenho original, para adaptações e mudanças ao longo do trajeto de modo que, em boa parte das vezes, o desenho final é bastante diverso daquele originalmente proposto. Mesmo assim, delinear a avaliação é essencial para que os recursos sejam otimizados, os resultados alcançados e as limitações conhecidas, na busca por sua qualidade. As três etapas de delineamento são a político-conceitual, a técnica e a administrativo-legal.

A etapa político-conceitual envolve uma discussão ampla e, muitas vezes, demorada na qual aqueles que decidiram por uma avaliação (os contratantes ou seus fomentadores, como os agentes financiadores) discutem por que e para que avaliar. Como as políticas públicas e programas sociais, de modo geral, têm objetivos muito amplos (LIPSKY, 1980), a discussão do “para que avaliar” tende inicialmente a ser vaga (embora não devesse sê-lo) ou improvisada (VIANNA, 1998), enquanto a justificativa da avaliação passa pela necessidade de “perseguir a qualidade” do seu objeto. Em vários casos, aqueles que contratam uma avaliação o fazem apenas porque “todo

20 Para conhecer um relato sobre o delineamento da avaliação de sistema, ler VIANNA, 2000a.

mundo avalia" e para dar "transparência às ações". Cabe à equipe avaliadora propor e conduzir discussões até que sejam definidas a razão para o esforço avaliativo (o por quê) e as necessidades a serem atendidas pela avaliação (o para quê), o que permitirá, então, delimitar o escopo da avaliação - o quê ou quem vai ser avaliado (o que não é o mesmo que quem vai ser entrevistado/testado/envolvido na coleta de dados como fonte) e quem vai avaliar (se indivíduo ou equipe, se interno ou externo). Também faz parte da etapa conceitual-política a determinação de por quanto tempo a avaliação deverá ocorrer.

Como posto por Stake (2004), a maior parte dos avaliadores aspira a uma prática profissional que leve, de modo geral, quem quer que esteja avaliando um determinado objeto aos mesmos achados. Isso, entretanto, não é factível. Na avaliação da qualidade e dos resultados de um programa (ou de uma intervenção ou ainda de uma política pública), não há uma única realidade a ser capturada: esse é um construto social e as pessoas têm posições sobre ele muitas vezes divergentes.

Respeitado o caráter político da avaliação, idealmente, esta etapa envolve não só os contratantes, mas principalmente os demais *stakeholders*, dentre os quais aqueles responsáveis pela implementação do objeto sob avaliação. A identificação desses *stakeholders* e, em paralelo, de outros usuários das informações oriundas da avaliação tem impacto no seu delineamento (inclusive pela definição do nível de desagregação do dado a coletar e da informação final) e, consequentemente, na concretização dos usos esperados. As perguntas que norteiam essa discussão são: qual o grau de autonomia do *stakeholder* / usuário da avaliação? O quanto ele pode mudar na realidade a partir dos resultados obtidos? O quanto ele faz parte do processo decisório maior? O uso dos resultados da avaliação está diretamente relacionado a tal grau de autonomia e às possibilidades de mudanças em cada contexto[21]. Considerando-se a ampliação do conceito de avaliação para incluir as dimensões de mérito e valor, pode-se dizer que essa etapa busca fundamentalmente os elementos voltados para o valor do objeto sob avaliação.

Uma vez definidos os elementos político-conceituais (quê, o por quê, o para quê, o por quanto tempo, e os "quem" da avaliação), a etapa técnica trata dos elementos metodológicos, ou o como. Há um vasto leque de abordagens técnico-metodológicas disponível e é importante identificar a que melhor atende – isoladamente ou fazendo parte de um *mix* – as definições conceituais e políticas. Algumas questões precisam ser respondidas: que dados, considerados relevantes na etapa anterior, serão buscados para informar sobre o objeto a ser avaliado? De que maneira essa

21 Patton (2003:39-62) propõe o incentivo de determinados usos para usuários definidos (*intended uses by intended users*), entre outras características, pelo grau de autonomia e poder decisório desses usuários.

definição será feita? (participativa ou técnica?) Um exemplo: se, na etapa conceitual, definiu-se que a avaliação deve informar sobre o nível de aprendizagem do alunado da Educação Básica na disciplina Matemática, na etapa técnica deve-se discutir como essa "aprendizagem em Matemática" será traduzida em elementos passíveis de avaliação. Esses elementos comporão um quadro operacional ou, no jargão da avaliação, sua matriz de referência, além, obviamente, de permitirem a determinação dos níveis de proficiência que vão caracterizar tal aprendizagem.

A etapa técnica relaciona-se aos elementos de mérito do objeto. Nela, além daquelas já listadas, devem ser respondidas as questões: Com que freqüência esses dados precisam ser coletados? Qual a melhor forma de acessá-los, tratá-los e analisá-los? Qual o nível de sigilo necessário? Qual o nível de controle na coleta, de modo a evitar desvios?

Há três elementos críticos que impactam as respostas às perguntas acima. O primeiro deles é tempo: quando os *stakeholders* precisam dos resultados? Quando se dá a tomada de decisão? De nada adianta os resultados avaliativos chegarem ao decisor após esse período. Essa limitação tem um grande impacto na freqüência de coleta de dados e na escolha do instrumento de coleta e das ferramentas para tratamento e análise de dados (chegando a interferir nos níveis de confiança). Como posto por Lawrenz, Gullickson e Toal (2007:287, tradução deste autor) "com freqüência, conseguir a informação para o *stakeholder* em tempo hábil tem algum custo em termos de precisão". A avaliação é definida pela perecibilidade: caso seus produtos não sejam entregues em tempo para o julgamento de valor (ou para a tomada de decisões), diferentemente da pesquisa (como visto na subseção anterior), perdem sua utilidade.

O segundo elemento crítico é o custo: delineamentos avaliativos que prevêem comparabilidade ao longo do tempo, informações muito desagregadas ou coletas censitárias, por exemplo, ainda que impactem positivamente a qualidade da avaliação e contribuam para a solidez de seus resultados, são muito mais caras que abordagens transversais, informações agregadas ou coletas amostrais. Visando à redução de custos, é importante, por exemplo, um levantamento prévio sobre informações já existentes, com possibilidade de utilização como dados secundários, com provável economia para a coleta de novos dados. Em outras situações, os dados secundários ou não foram coletados com o rigor necessário, ou têm nível de desagregação menor que o demandado ou ainda estão tão distantes do objeto sob a avaliação que a melhor abordagem é a coleta de dados novos. Além disso, é menos caro o delineamento avaliativo com dados existentes que aqueles que prevêem a criação de dados, usando-se a categorização proposta por Laville e Dionne (2001).

O quanto o contratante é capaz de sustentar o delineamento escolhido ao longo do tempo pretendido define, em grande parte, o desenho avaliativo. A etapa técnica compreende também as projeções de custo, dado o horizonte temporal previsto, de modo a evitar uma interrupção indesejada durante o processo ou o abortamento da própria política por falta de recursos.

O terceiro elemento crítico que restringe o delineamento da avaliação é a capacidade[22] institucional e a competência da equipe avaliadora. As abordagens propostas devem ser tais que a equipe existente (ou a contratar) seja capaz de conduzi-las, especialmente quando o contratante é esfera governamental, não competitivo para captação de talentos. Segundo Calmon (2005), a capacidade institucional depende de quatro elementos, a saber:

> - Processos que permitam a coleta e análise sistemática de informações, identificação de problemas e formulação de soluções;
> - Atores (individuais ou coletivos), que possuam conhecimento, aptidão, recursos materiais e motivação para atuar de forma eficaz;
> - Organizações apropriadas, transparentes e participativas que viabilizem e estruturem a ação coletiva desses atores;
> - Instituições entendidas como sendo as normas e regras (formais e informais) que incentivem a atuação eficaz dos atores e contribuam para a sustentabilidade política, econômica e social das suas ações. (CALMON, 2005:6-7).

Para equipes iniciantes, o delineamento da política de avaliação pode prever o desenvolvimento da competência avaliadora da equipe, por um lado, e a busca de sinergia entre os quatro elementos acima, no sentido de favorecer esse desenvolvimento.

A etapa técnica considera também a determinação de como serão divulgados/disseminados os resultados da avaliação e de como serão envolvidos os *stakeholders* nesse processo. Por fim, diante do número de dados a coletar, especialmente quando observado o horizonte temporal e a quantidade de dados a processar e a manter, na etapa técnica se identificam os suportes para tais dados (equipamentos, ferramentas, etc.) e as formas e políticas de armazenamento e acesso.

Os resultados da etapa técnica são, a partir dos elementos críticos e do panorama metodológico disponível, confrontados com as possibilidades de implementação e, em boa parte dos casos, são realinhados e redefinidos para que a política de avaliação seja implementada de modo preciso e válido[23].

---

22 Para definição de capacidade, optou-se pela proposta apresentada por Calmon (2005: 6): "habilidade de compreender e analisar uma determinada situação, identificar problemas, definir e implementar metas, objetivos e formular estratégias para ações futuras".

23 Diz-se de uma avaliação (ou um teste) que é válida quando é capaz de responder as perguntas para as quais foi delineada. Se um teste é aplicado a alunos de 2ª série para medir as competências e habilidades da 2ª série em

Melhor visualizados os elementos político-conceituais e os técnicos, é importante uma busca quanto às questões administrativo-legais que lhes são atinentes, para garantir as possibilidades de implementação do desenho proposto e para assegurar que o processo avaliativo ocorra dentro de limites estabelecidos pelas bases legais vigentes. Fazem parte dessa terceira fase a discussão das rubricas orçamentárias e de repasse de recursos ao longo do tempo, as possibilidades de contratação e modificação da equipe de avaliação, as formas de aquisição de equipamentos e *software*, a instalação da equipe, e, principalmente, um aprofundamento sobre a base legal que rege o objeto da avaliação e o levantamento das restrições de acesso aos dados existentes ou a determinadas fontes. Novamente, é freqüente, diante do panorama administrativo-legal, a restrição do delineamento proposto nas fases anteriores (em muitos casos, o orçamento será o elemento definidor final).

Todas as perguntas acima respondidas, é crucial que o planejamento – e posteriores implementação e avaliação – de uma avaliação em larga escala assegure sua qualidade. Mas, o que determina a qualidade da avaliação e, na mesma linha, de uma política de avaliação?

Associações de avaliação, de modo geral, representações de instituições avaliadoras, representações governamentais, e luminares no campo da avaliação vêm, ao longo dos últimos trinta e cinco anos, discutindo a qualidade da avaliação. De início, a discussão voltava-se para aspectos mais técnicos ou isolados. Mais recentemente, a definição da qualidade passa pelo atingimento de uma série de padrões propostos e validados pelas representações mencionadas antes. Em algumas abordagens, os critérios referem-se à experiência avaliativa; em outros casos, à conduta do avaliador. Na maior parte dos casos, não são específicas para políticas de avaliação, mas podem ser adequadas a elas. A subseção a seguir apresenta algumas dessas propostas no sentido de dar ao leitor o panorama sobre a qualidade da avaliação que, em última análise, forma a base para a matriz da meta-avaliação. Dessa matriz, foram buscados os conceitos-chave – Uso e Utilidade – do estudo em relato.

---

matemática, as questões de leitura não podem interferir nessa medição. Ou se um questionário é aplicado para se levantar a opinião de professores sobre uma determinada política, seus resultados não podem ser usados para a discussão sobre fatos relacionados a essa mesma política. Por essa razão, é fundamental que os objetivos da avaliação estejam definidos ANTES da sua implementação. Diz-se de uma avaliação que é precisa quando retrata de maneira acurada a realidade sobre a qual se debruça. Essa característica está relacionada à exatidão dos dados levantados. Entretanto, quando há pessoas envolvidas, é muito difícil obter-se uma medida exata, já que há uma série de fatores não totalmente controláveis interferindo na medição. Por essa razão, quando uma alta precisão não é possível, é imprescindível que o usuário da informação seja alertado sobre seus limites.

2.2.5 Padrões e critérios que indicam a qualidade da avaliação

Historicamente, padrões e critérios de qualidade para avaliação têm sido definidos por associações e por indivíduos expoentes no campo avaliativo e têm como foco ora o delineamento avaliativo, ora o comportamento do avaliador. Por vezes, tais padrões são "importados" daquilo que é considerado qualidade em pesquisa científica (especialmente quando observados os aspectos mais técnicos, como os relativos aos instrumentos de coleta e à forma de tratamento e de análise de dados) [24]. Isso se dá pela já discutida proximidade de uma e outra e, fundamentalmente, porque o campo da avaliação é mais novo que o da pesquisa científica, não havendo ainda uma consolidação da avaliação como campo de ação (profissão) ou mesmo área de conhecimento[25]. Nessa subseção, padrões e critérios são utilizados como termos equivalentes para sinalizar características da avaliação que, se presentes na experiência, apontam para sua qualidade.

Para a discussão sobre padrões de qualidade da avaliação, a seguir são listadas algumas associações e os padrões por elas adotados, o posicionamento adotado por pesquisadores e avaliadores que apresentam suas críticas e suas propostas de outros padrões ou critérios. Inicialmente, foram relacionados padrões que definem uma avaliação de qualidade; em seguida, foram dispostos padrões e critérios sobre a conduta do avaliador. No Brasil, na ausência de padrões validados nacionalmente[26], apresentam-se critérios dispostos em políticas como, por exemplo, o SINAES.

Em termos de padrões de qualidade para as experiências avaliativas, os EUA têm sido referência mundial. Em 1974, três grandes associações norte-americanas (APA, AERA e NCME) propuseram uma força-tarefa para revisar os padrões de seus testes. Dessa força tarefa surgiu o *Joint Committee on Standards for Educational Evaluation* (JCSEE), criado para, a partir de uma visão mais abrangente que aquela então vigente, propor novos padrões para avaliações. Em 1981, o JCSEE publicou padrões para avaliações de programas, projetos e materiais educacionais. No entendimento de que deveria haver uma equipe permanente voltada para a manutenção e melhoria dos padrões, o Comitê se institucionalizou, transformando-se em uma organização sem fins lucrativos que, em 1989, torna-se certificada, submetendo seus padrões ao *American National Standards Institute* (ANSI). Em 1994, o JCSEE publica então um conjunto validado de trinta

24 Para aprofundar a leitura sobre padrões, consultar HARTZ, 2006.
25 Para uma leitura sobre a consolidação da área de avaliação, consultar o terceiro capítulo de WORTHEN; SANDERS; FITZPATRICK, 2005:73-96
26 A ABAVE – Associação Brasileira de Avaliação Educacional – até o momento não apresentou qualquer projeto neste sentido, por exemplo.

padrões distribuídos em quatro categorias[27] (Precisão – 12 padrões, Viabilidade - 03, Propriedade – 08, e Utilidade 07) que, desde então, têm influenciado a criação de padrões por muitas associações e entidades de avaliação pelo mundo afora e que dão base para várias experiências de meta-avaliação. O Apêndice 02 apresenta uma figura que ilustra os trinta padrões nas quatro categorias do JCSEE.

A categoria Precisão abrange os critérios que tratam da qualidade da implementação da avaliação, do desenho dos instrumentos, da coleta, tratamento e análise de dados, entre outros. Ou seja, concentra-se nas questões técnicas da avaliação, sua confiabilidade, validade e precisão, os registros feitos dos procedimentos adotados, a sistematização da avaliação, a análise do contexto, dentre outros aspectos. Nessa categoria, um dos padrões é a meta-avaliação (interessantemente relacionada com técnica). A categoria Propriedade volta-se para os *stakeholders* e para seu atendimento, para a qualidade das relações estabelecidas, e para questões éticas e legais. A categoria Viabilidade analisa os aspectos de factibilidade da avaliação, incluindo capacidade da equipe de avaliadores e relação custo/efetividade. A categoria Utilidade, por fim, mistura indicadores de disseminação dos resultados com o levantamento do perfil dos usuários da avaliação. Tal categoria constitui a real contribuição do JCSEE para o entendimento da qualidade da avaliação. Na atualidade, para alguns pesquisadores como Letichevsky *et alii* (2007), essa é a categoria que define essencialmente se a avaliação primária tem ou não qualidade. O texto completo com os padrões está disponível em http://www.wmich.edu/evalctr/jc/. Organizados em uma listagem de sentenças relativamente curtas, os padrões do JCSEE são facilmente convertidos em *checklists*, para verificação do seu cumprimento pelas avaliações sob análise, como feito por Stufflebeam.

Ainda que facilmente transmitidos, a utilização desses padrões americanos em outros cenários nacionais não é tão simples. Por exemplo, a experiência suíça de aplicação dos padrões para análise de um conjunto de 15 experiências avaliativas, relatada por Widmer (2005), mostra que o primeiro obstáculo é a tradução do inglês (no caso, a tradução para o alemão) e, em seguida, os ajustes para um contexto de avaliação local (no caso suíço, também para uma situação não educacional).

Widmer (2005) critica o grande número de critérios e sugere que seja reduzido. Há uma discussão, inclusive no Brasil[28], que abrange, por um lado, a necessidade de informar sobre todos

27 Em inglês, *accuracy*, *feasibility*, *propriety*, *and utility*.
28 Sobre a redução de critérios na ficha de avaliação da CAPES, ver VERHINE, 2008.

os aspectos que são associados à qualidade (seja da avaliação, seja do objeto) e, por outro, da dificuldade em consolidar um veredicto final, que aponte os pontos fortes e fracos, mas sem dissipação. Como reduzir o número de critérios sem perder a qualidade da meta-avaliação? Aliás, sem deixar de informar o avaliador sobre as questões que afetam seu processo avaliativo? É possível que o número de critérios esteja relacionado à necessidade de ajudar os avaliandos a contribuir para a melhoria do seu objeto e também à necessidade de formar os avaliadores pelas meta-avaliações.

No caso da experiência suíça com os padrões do JCSEE, outras críticas são relativas ao fato de que os padrões se sobrepõem (e entram em conflito), em alguns casos dificultando o posicionamento do avaliador, ou, dito de outra forma, que um mesmo aspecto da avaliação é coberto por mais de um padrão. Tal experiência demonstra também que nem todos os critérios são alocados em um mesmo nível analítico. Tais problemas, segundo Widmer (2005), tornam quase impossível a aplicação direta e "dura" dos padrões, o que será visto de maneira mais detalhada na subseção que trata dos indicadores da categoria Utilidade.

De maneira mais pragmática, pode-se considerar a abordagem utilizada por Chelimsky (1983 *apud* REBOLLOSO *et alii*, 2002:15), de número bem menor de indicadores, na qual são diferenciadas a adequação técnica (*technical appropriateness*) e a utilidade da avaliação. A primeira busca levantar se o delineamento é adequado para as necessidades dos usuários. Essa dimensão é relativa (definida, inclusive, como a diferença entre o custo de um delineamento e outro, dadas as necessidades dos *stakeholders*) e demanda um equilíbrio entre diversos elementos do delineamento: adequação de planejamento, de execução, a existência de outras opções viáveis e a ausência de erros conceituais. A utilidade volta-se para a análise da avaliação a partir de quatro componentes: relevância dos dados em vista das necessidades de informação, pontualidade (no sentido de perecibilidade), apresentação do relatório de avaliação, e os usos dele decorrentes. Todas as medidas de utilidade enfatizam a implementação de aspectos do delineamento negociados pelo cliente e pelo monitoramento de sua satisfação e do uso.

Uma outra crítica aos padrões do JCSEE, além das questões observadas por Widmer (2005), é que não trazem consigo os níveis de qualidade de um processo avaliativo, quando observados em conjunto. Em outras palavras, é possível observar se uma experiência avaliativa cumpre ou não um determinado conjunto de padrões e que não atinge outros tantos, mas não há um julgamento global. Essa avaliação seria ou não considerada uma experiência de qualidade? Uma discussão sobre a escala de avaliação foi feita por Letichevsky *et alii* (2005 e 2007). Sem desmerecer tal

trabalho, a categoria Precisão focalizada nesse trabalho, no entanto, é justamente aquela que mais aproxima a avaliação da pesquisa, havendo, nesse sentido, uma base teórica consolidada que pode ajudar o grupo a definir a escala. Para as dimensões Propriedade e, principalmente, Utilidade, a definição da escala parece mais complexa.

Saindo da abordagem do JCSEE, Gingsburg e Rhett (2003) nos lembram que, assim como na pesquisa científica, são as questões de avaliação que determinam a melhor metodologia a utilizar. E referem a um aumento da demanda, por parte dos formuladores de política e do Congresso americano, pelo rigor no delineamento da avaliação, o que permite aos avaliadores adotar metodologias mais adequadas, por um lado, e por outro ter seu trabalho validado e bem utilizado. Esses autores citam o *Institute for Education Sciences* que, por meio de bases legais determinadas pelo Congresso americano em 2002, determina, com viés nitidamente quantitativo, que uma avaliação científica válida (vale a pena ressaltar o uso da palavra científica) deve aderir aos mais altos padrões de qualidade existentes no que diz respeito ao modelo de investigação e de análise estatística; empregar delineamentos experimentais com utilização de amostragem randômica, quando possível, e outros desenhos que favoreçam a identificação de inferências causais no caso da amostragem randômica não ser possível; e estudar a implementação de programa por meio de uma combinação de métodos cientificamente válidos e confiáveis (GINSBURG; RHETT, 2003). Na contrapartida, Stake (2004) chama atenção para o fato de que dois avaliadores competentes, avaliando um mesmo objeto, dificilmente chegarão aos mesmos resultados e, por isso, o uso da meta-avaliação é essencial, ainda que não elimine incertezas nos achados da avaliação.

Já na Europa, se alguns países ainda discutem seus padrões para a avaliação (como no caso mencionado por Widmer em relação à Suíça), em 2003, na sua *Charte de l´évaluation des politiques publiques et des programmes publics* (texto disponível em www.sfe.asso.fr/docs/*site*/charte/sfe_charte_2003.pdf), a *Société Française de l´évaluation* definiu e divulgou, para a avaliação de políticas públicas e a partir do entendimento de que ela precisa ser voltada para o bem geral do conjunto de cidadãos, seis princípios: pluralidade, distanciamento, competência, respeito às pessoas, transparência e responsabilidade, em um posicionamento ético que a aproxima mais da *American Evaluation Association* (ver a seguir) que do JCSEE. Para os membros dessa Sociedade, a avaliação deve ser exercida a partir de um quadro institucional explícito. Ainda em 2003, o Reino Unido lançou seu Guia de Boas Práticas de Avaliação, também se distanciando um pouco da proposta do JCSEE.

Na América Latina, o PREAL (*Programa de Promoción de la Reforma Educativa en América Latina y el Caribe*) divulgou um conjunto de 10 recomendações para as avaliações educativas (disponível em http://www.preal.org/Grupo3.asp?Id_Noticia=156&Id_Grupo=3), relacionados a seguir:

> 1. A avaliação deve ser concebida como um elemento articulado em um conjunto mais amplo de ações e políticas educativas.
> 2. A avaliação deve contemplar um processo de reflexão coletiva sobre o estado da educação e sobre os caminhos para melhorá-la.
> 3. A avaliação deve estar a serviço do desenvolvimento de um sentido de responsabilidade compartilhada pela educação como bem público.
> 4. Os sistemas de avaliação necessitam ampliar progressivamente o leque de fins educativos que são objeto da avaliação.
> 5. É fundamental desenhar avaliações do progresso dos alunos (estudos longitudinais).
> 6. Um sistema de avaliação é um projeto de longo prazo, que requer um compromisso de Estado e um planejamento cuidadoso do desenho do sistema.
> 7. Um bom sistema de avaliação demanda investimento.
> 8. O sistema de avaliação deve ser fundamentado em uma sólida postura de transparência.
> 9. Os ministérios de educação devem assumir um compromisso sério com os resultados da avaliação.
> 10. Os sistemas de avaliação devem ser objeto de avaliação periódica. (RAVELA *et alii*, 2008).

Nesse último caso, o PREAL segue os padrões do JCSEE, ainda que mais voltados para a política pública educacional, e recomenda, no seu 10º princípio, a meta-avaliação como condição de qualidade para a avaliação. Os demais parecem ser direcionados aos gestores de sistemas educacionais, responsáveis pelas "encomendas de avaliação", e são facilmente relacionados às três etapas de planejamento de uma avaliação, discutidas na subseção anterior.

Organismos internacionais, como o Comitê de Assistência ao Desenvolvimento da OCDE[29] e sua Rede de Avaliação, o Banco Mundial, e a UNESCO têm também seus critérios para a definição do que seja uma avaliação de qualidade. No caso do referido Comitê, por exemplo, há uma orientação geral, a partir de pilares centrais, para as avaliações independentes (tanto o processo quanto o produto) de modo que possam, por um lado, auxiliar na definição de expectativas dos países membros quanto às avaliações e, por outro lado, oferecer um guia de boas práticas de avaliação para melhor servir às intervenções desenvolvimentistas, de maneira a contribuir para uma abordagem harmônica da avaliação alinhada aos princípios da *Paris Declaration on Aid Effectiveness*. O Banco Mundial disponibiliza, em seu *site*, tutoriais para estatística, por exemplo, e informações sobre o que considera as boas práticas da avaliação.

A UNESCO (2007), em seu Manual da Avaliação (cujo público-alvo são as organizações ligadas às Nações Unidas), divulga os padrões estabelecidos pela UNEG – Grupo de Avaliação das Nações Unidas (2005) – para a qualidade de avaliação. Eles estão divididos em 04 categorias: (1) Arcabouço institucional e gestão da função avaliativa; 2) Competência e ética; 3) Condução da avaliação (na qual está incluída a etapa de desenho da experiência avaliativa); e 4) Relatórios da Avaliação. Chama atenção o destaque dado pelo Grupo e pela UNESCO aos relatórios e à comunicação, especialmente quando na categoria anterior (Condução), há uma seção inteira para a disseminação dos resultados. Não há, no entanto, uma seção para os usos da avaliação, ainda que a UNESCO espere, como *follow-up*, uma manifestação de governos e entidades quanto aos elementos apontados pelos relatórios de avaliação. No caso da UNEG, cada padrão é seguido por uma recomendação, que o detalha e que orienta o avaliador a seguir a norma. É também interessante perceber o destaque à gestão da avaliação, o que torna a recomendação da UNEG mais aplicável em termos de experiências em larga escala. Em relação à meta-avaliação, um dos padrões considera uma revisão por pares ou por um grupo de referência particularmente útil. Para a UNESCO, são princípios-chave para a avaliação: independência, imparcialidade, intencionalidade, perecibilidade, transparência, competência, ética, e qualidade (relacionada ao cumprimento dos padrões da UNEG).

No Brasil, há dois tipos de padrão: no primeiro caso, restrito, encontram-se aqueles oriundos de encomendas específicas, como o exemplo citado por Hartz (2006) ao discorrer sobre o termo de referência proposto no edital de contratação da meta-avaliação externa para os Estudos de Linha de Base (ELB/PROESF) da área de Saúde. No referido termo consta a definição de um modelo a utilizar e nele são incorporados “dimensões analíticas inter-relacionadas à implantação e ao impacto do programa, à organização e integralidade do cuidado, ao desempenho dos sistemas locais e ao contexto político” (HARTZ, 2006:735). Embora não sejam padrões gerais do campo de avaliação, pelo menos a expectativa dos *stakeholders* contratantes é bem delineada por edital (encomenda avaliativa ou meta-avaliativa).

A segunda forma de uma conceituação da qualidade para a avaliação pode ser vista nos textos de apresentação de políticas como, por exemplo, o SINAES, nos quais são apresentados os princípios que regem aquela avaliação (ou sistema de avaliação). Para os formuladores da política de avaliação da educação superior, hoje vigente no Brasil, a avaliação é uma prática social com fins educativos, o que pressupõe um caráter formativo. São também princípios o respeito à diversidade

---

29 Para maiores informações sobre a DAC Evaluation Network, consultar o site www.oecd.org/dac/evaluationnetwork.

e à individualidade, a globalidade, a legitimidade e a continuidade (INEP, 2007). Nos textos institucionais do SINAES ou na sua base legal não há, no entanto, qualquer proposta de detalhamento de parâmetros de meta-avaliação da política, sendo mais um posicionamento político-ideológico que uma definição técnica da qualidade da experiência avaliativa. O mesmo ocorre, como pode ser visto a seguir, com posições defendidas individualmente por pesquisadores e acadêmicos.

Sguissardi (1995), ao levantar as questões associadas às políticas de avaliação no Brasil e seu relacionamento com o que chamou de "filosofia da qualidade total", propôs alguns padrões para as experiências avaliativas com foco no ensino superior. Para ele, "para avaliar propostas de avaliação no ensino superior, deve-se começar pelas respostas a questões tais como: a) o que de fato fundamenta e justifica a avaliação"? b) para que serve a avaliação? c) quais as principais questões que têm sido levantadas diante das propostas de avaliação de iniciativa oficial e mais recentemente também de iniciativa de organismos ligados a entidades empresariais privadas? (SGUISSARDI, 1995:562). Esse mesmo autor cita M. L. Cardoso ao propor um conjunto de critérios para meta-avaliação, transformados em perguntas e recomendações meta-avaliativas para análise de experiências de avaliação institucional da educação superior:

> - Uma proposta ou um processo de avaliação é função de um projeto de desenvolvimento da sociedade. Isto pressuposto, tornam-se obrigatórias questões como: De que competência se trata? Competência para quê? Competência para integrar-se e servir a (ou questionar) que tipo de sociedade? - Uma proposta ou um processo de avaliação traz em si e contribui para implantar ou fortalecer um dado padrão de política educacional e/ou de universidade. E isto precede à discussão imediatamente técnica que pergunta quem e como avalia. Antes devem ser formuladas questões como: Qual universidade? Universidade para quê?
> - Admitida a necessidade da avaliação, esse processo deveria abranger todo o sistema escolar e todas as suas atividades, voltado para a elevação da qualidade da educação nacional em todos os graus da rede pública e privada.
> - A avaliação deve estender-se à totalidade das atividades da instituição objeto desse processo.
> - A avaliação deve ser ampla, global e se iniciar com a universidade enquanto instituição: cada universidade tem um perfil, tem uma história. É preciso identificar esse perfil, reconstituir essa história, para avaliar o papel que essa universidade específica tem desempenhado historicamente na sociedade e diante do desenvolvimento da ciência e colocá-lo em discussão, especialmente para definir se é isso mesmo que a comunidade universitária (e também a comunidade em geral) quer (em) para essa universidade neste momento e no futuro próximo.
> - Nesse tipo de avaliação institucional global é básica a análise das verbas com as quais trabalha a universidade: quanto recebe, de que fonte(s) e de que forma (global, parcelada); por outro lado, como são distribuídos internamente esses recursos (quem decide, critérios, setores contemplados, regularidade).
> - Para que a avaliação institucional global seja completa é preciso analisar e avaliar o processo decisório no interior da universidade: Quem decide e como são tomadas as decisões (se existem ou não mecanismos públicos de controle

> sobre o processo de tomada de decisões)? Qual o grau de abertura e de flexibilidade da estrutura de poder dentro da universidade em relação à sociedade? Que lugar ocupam e qual a importância que os movimentos sociais organizados têm nas deliberações?
> - A avaliação do desempenho acadêmico, como parte da avaliação institucional global, deve incluir todas as atividades de ensino (professores, monitores, estudantes), pesquisa (pesquisadores, técnicos) e administração (em todos os níveis, envolvendo desde os servidores que ocupam os cargos mais simples até os dirigentes máximos das instituições). O processo deve abranger, por um lado, unidades e departamentos e, por outro, os cursos, e só então o professor e o pesquisador individuais.
> - Um processo de avaliação institucional global dessa natureza deve ser explícito e claro e pressupõe um projeto de universidade, que seja legítimo pela forma de sua construção, resultante de discussão aberta, ampla e democrática, livre e coletiva, e oposta a qualquer imposição de tipo tecnocrático.
> - O centro desse projeto de universidade é uma política acadêmica, que compreende fundamentalmente uma política de ensino, uma política de pesquisa, uma política de extensão e uma política administrativa presumivelmente associadas. (CARDOSO, 1991, p.22-23 *apud* SGUISSARDI, 1995:570).

Como pode ser visto, os parâmetros apresentados são por demais abrangentes e não se prestam a uma análise meta-avaliativa mais detalhada, embora contribuam para o delineamento da experiência de avaliação. Discute-se ser possível, por um lado, o incentivo a experiências de auto-avaliação, em respeito ao contexto e história do objeto avaliado; por outro lado, é importante uma análise referenciada em experiências outras, que permitam ao objeto avaliado situar-se no panorama mais geral.

No Brasil e no mundo, avaliadores de renome, individualmente, dedicaram-se e continuam a dedicar-se a identificar padrões que determinem qualidade do processo avaliativo. Por exemplo, Chen (1988, 1990 *apud* REBOLLOSO *et alii*, 2002:14) propôs uma síntese de critérios definidores de qualidade para a avaliação: objetividade, que pode ser vinculada ao critério de confiabilidade, confiança, generalização e efetividade, que inclui valores como perecibilidade (*timeliness*), relevância e amplitude de conseqüências. No Brasil, Vianna (1998; 2001; 2003) tem discutido aspectos da avaliação em larga escala no sentido de alertar avaliadores sobre seus problemas, de modo a antecipá-los e favorecer sua solução. Dentre os mais citados entre tais autores estão Scriven e Stufflebeam, ambos envolvidos na base de formação do JCSEE. Para Stufflebeam (1974), por exemplo, as avaliações deveriam atender aos padrões de adequação técnica, utilidade e custo/efetividade. Esses padrões foram posteriormente definidos na forma de onze critérios: validade interna, validade externa, confiabilidade, objetividade, relevância, escopo, importância, credibilidade, pontualidade, penetração, e custo/eficiência. Como se pode perceber, em 1974, ainda não estava lançada, de modo explícito, a discussão sobre os aspectos relacionados ao mérito e ao valor do objeto avaliado ou mesmo da própria avaliação, que surge em 1987. Como

já mencionado, o mérito está diretamente relacionado à qualidade técnica e aos aspectos instrumentais de um processo avaliativo: a avaliação fez bem aquilo que se dispôs a fazer (DAVOK, 2006). Já o valor compreende os aspectos que ligam a avaliação ao atendimento das necessidades dos seus *stakeholders* (tanto em planejamento quanto no uso). O conceito de valor confunde-se, em alguns aspectos, com o de efetividade, se pensarmos como Sander, para quem efetividade, correspondente ao termo em inglês *responsiveness*, é um critério político voltado para o atendimento, pela administração pública, de demandas sociais (SANDER, 1995).

O objetivo dessa subseção é introduzir o debate sobre critérios de qualidade para a avaliação, de modo a que se possa, mais adiante, discutir sua aplicação pela meta-avaliação, mas não se pretende aqui exauri-los. Por essa razão, os parágrafos a seguir deixam o foco na qualidade da avaliação para dedicarem-se aos critérios de qualidade da conduta do avaliador, a partir da proposta da *American Evaluation Association* -AEA (1994).

Os cinco princípios da AEA, publicados em 1994 e revisados dez anos depois, em 2004, compõem uma recomendação de conduta para os avaliadores, em lugar de orientar as escolhas para a avaliação. O trecho introdutório dessa recomendação, disponível na íntegra em www.eval.org, lembra que os princípios se sobrepõem de muitas maneiras e, eventualmente, podem ser conflitantes, o que demanda um posicionamento dos avaliadores que, de resto, merece consulta a outros profissionais da área quando não há certeza sobre como proceder. Nesse sentido, diferentemente dos padrões do JCSEE, é mais difícil uma derivação desse documento em *checklists*. Os cinco princípios são colocações norteadoras de conduta e não elementos de verificação de qualidade. São eles: I) investigação sistemática; II) competência; III) integridade e honestidade; IV) respeito às pessoas; e V) responsabilidade com o bem-estar geral. De acordo com Hartz (2006:734), esta última recomendação foi registrada pela primeira vez como uma diretriz formal de conduta nesse documento da AEA.

Cada princípio é desdobrado em um número de recomendações, no total de 25, como se pode ver a seguir. É interessante observar que a Associação justifica ou exemplifica, na medida do possível, cada uma das recomendações e que, de maneira geral, conduz à reflexão do avaliador na escolha do caminho a trilhar, em lugar de determinar a trilha correta[30].

30 Na tradução dos *Guiding principles for evaluators* feita para a elaboração do presente texto, este autor optou por simplificar o texto original da *American Evaluation Association*, concentrando-se na recomendação e não traduzindo

O princípio da investigação sistemática pressupõe que o avaliador: 1) utilize os mais altos padrões dentro da abordagem escolhida para assegurar precisão e credibilidade pela informação da avaliação; 2) discuta com seu cliente sobre pontos fortes e fracos das várias perguntas de avaliação e das abordagens escolhidas para respondê-las; e 3) comunique suas escolhas de métodos, abordagens e resultados com suficiente grau de detalhamento – inclusive das limitações – e contextualização que permita ao receptor entender, interpretar e criticar o trabalho.

É importante perceber que o reconhecimento das limitações não é algo posto para os formuladores de política pública, como o é para pesquisadores e a academia em geral. A negociação entre formuladores e avaliadores, portanto, é bastante sensível, mas as limitações devem ser colocadas tanto no delineamento, quanto ao final, nos relatórios de avaliação, o que tende a minimizar o mau uso das informações.

O princípio da competência recomenda que o avaliador: 1) adquira (ou tenha na equipe de avaliadores) os conhecimentos, as habilidades e a experiência adequadas no cumprimento das atividades previstas na avaliação; 2) demonstre competência cultural[31], de modo a assegurar reconhecimento, interpretação acurada e respeito à diversidade; 3) conduza sua prática nos limites da sua capacitação e competência profissionais e que recuse trabalhos que estejam fora desses limites; e 4) busque continuamente manter e melhorar as competências adquiridas, de modo a oferecer o melhor possível em sua prática.

No princípio que rege integridade e honestidade, as recomendações são no sentido de que o avaliador: 1) assuma a liderança e negocie de maneira honesta com clientes e *stakeholders* os custos, as atividades, as limitações, o escopo possível para futuros resultados e os usos de dados em uma experiência avaliativa; 2) revele, antes do início de um processo, quaisquer relações ou questões que possam configurar-se conflito de interesse e, no caso de seguir com a avaliação, dispor sobre o conflito nos relatórios finais; 3) registre todas as mudanças feitas no projeto original e as razões para adotá-las (no caso de grande impacto, as mudanças devem ser informadas aos clientes e *stakeholders* antes de implementadas); 4) seja explícito sobre seus próprios interesses e valores na condução e nos produtos da avaliação, bem como os dos clientes e

---

as justificativas. O texto integral dos *Guiding principles* pode ser encontrado no *site* da AEA (www.aea.org).e em www.eval.org.

31 De acordo com a *American Evaluation Association*, a competência cultural deve ser refletida, no avaliador, em sua conscientização das próprias crenças culturais, no entendimento das visões de mundo dos participantes oriundos de culturas diversas e no uso de estratégias e habilidades adequadas para o trabalho em grupos culturalmente diferentes. A diversidade pode ser em termos de raça, etnia, gênero, religião, origens sócio-econômicas, ou outros fatores presentes ao contexto sob avaliação (AMERICAN EVALUATION ASSOCIATION, 1994).

stakeholders; 5) dificulte a má utilização ou má interpretação de seus procedimentos, dados ou achados por si mesmo ou, no limite das possibilidades, pelos outros; 6) comunique suas preocupações e razões quanto a procedimentos ou atividades que facilitem a produção de informação ou conclusão enganosa e, no caso de o cliente não dissipar essas preocupações e de ser impossível rejeitar a avaliação, consulte pares e *stakeholders* na busca de solução para o conflito; e 7) revele todas as fontes financiadoras e de interesse em cada experiência de avaliação. Nesse momento, volta-se à discussão sobre o que se constitui um impedimento para que o avaliador aceite uma avaliação: viés pessoal, como, por exemplo, o envolvimento do avaliador com a "causa" sob avaliação; conflito ideológico; quebra de harmonia com a equipe, dentre outros. Stake chama atenção para o fato de que o princípio pede a declaração de conflitos, mas nada diz sobre as confluências de interesses que, de um modo ou outro, podem vir a afetar o delineamento avaliativo (STAKE, 2004).

O quarto princípio traz recomendações que envolvem o respeito por pessoas, sejam elas as fontes de dados, os participantes do programa avaliado, os clientes ou outros *stakeholders*. Esse princípio reza que o avaliador deva: 1) ter uma compreensão abrangente dos elementos contextuais da avaliação; 2) respeitar padrões, regulamentações e a ética profissional no sentido de que seu trabalho não implique danos, riscos e fardos para aqueles que dele participam[32]; 3) maximizar os benefícios e reduzir possíveis danos decorrentes das avaliações (em especial, dos resultados negativos), desde que isso não comprometa a integridade dos achados; 4) comunicar os resultados de um modo que respeite inequivocamente a dignidade e o valor dos *stakeholders*; 5) sempre que possível, favorecer a eqüidade social na avaliação, de forma a que aqueles que tenham contribuído com ela possam dela beneficiar-se; 6) entender e respeitar as diferenças entre os participantes e levá-las em consideração quando do planejamento, implementação, análise e comunicação dos resultados da avaliação.

O quinto e último princípio diz respeito às responsabilidades para o bem estar geral e público. As cinco recomendações que compõem esse princípio determinam que o avaliador: 1) considere as perspectivas e interesses de todos os *stakeholders* quando planejando ou comunicando a avaliação; 2) tenha em mente não apenas as operações e resultados imediatos, mas também as implicações e efeitos colaterais potenciais; 3) garanta o acesso de todos os *stakeholders* relevantes às informações de modo a respeitar as pessoas e a honrar compromissos assumidos sobre sigilo; ao buscar esse acesso, comunicar de maneira clara e simples de modo a que os clientes e

32 Inclusive com a assinatura de termos de consentimento informado sobre o escopo e sobre limites de sigilo da avaliação.

stakeholders possam facilmente compreender processos e produtos da avaliação; 4) promova o equilíbrio entre as demandas dos clientes e dos demais, buscando identificar, discutir e resolver possíveis conflitos entre eles; 5) obrigue-se com o interesse e o bem públicos, muitas vezes precisando ir além dos interesses particulares de grupos específicos envolvidos com a avaliação.

O quinto princípio, muito próximo da categoria Utilidade do JCSEE, deve ser visto com algum cuidado já que, especialmente no subitem 5, ele pode induzir uma confusão entre avaliação e intervenção. Ir além dos interesses específicos em uma determinada experiência avaliativa pode ou não ser factível e depende tanto do avaliador quanto (e principalmente) dos responsáveis pela encomenda avaliativa. Há também uma provável sobrecarga de trabalho que nem sempre o demandante deseja arcar nesse contexto. Cabe ao avaliador, pelo menos, propor e, no momento da elaboração dos relatórios finais, recortar exatamente o objeto avaliado.

Da mesma maneira que ocorre quanto à discussão sobre a qualidade da experiência avaliativa, também quanto à conduta do avaliador há propostas individuais, não apresentadas por associações. Por exemplo, em 1997, Shulha e Cousins (1997) recomendavam precaução contra situações que pudessem facilmente levar ao mau uso da avaliação, sintetizando-a em três recomendações: busca de verificações independentes dos seus processos avaliativos, envolvimento em revisões metodológicas e a consulta a códigos de boas práticas existentes.

Todo o panorama apresentado nessa subseção indica uma demanda por qualidade – política e técnica – das avaliações, apresentada por indivíduos ou por associações. No presente trabalho, esta subseção do marco teórico é importante por favorecer a noção de que não há um padrão único e que muito precisa ser feito em termos da conceituação da qualidade seja da avaliação e da política de avaliação, seja da conduta do avaliador. Além disso, propor critérios, princípios, padrões de qualidade não implica a verificação de seu cumprimento. É essa a lacuna preenchida pelas meta-avaliações (BUSTELO, 2006). O marco teórico, a partir desse momento, volta-se para a conceituação de meta-avaliação e posterior discussão sobre como concretizá-la para, por fim, restringir seu foco às categorias Utilidade e Uso, bases do quadro de análise da presente pesquisa.

2.3 Meta-avaliação

A presente pesquisa não foi concebida como uma meta-avaliação e não relacionou, dentre seus objetivos, julgar a política escolhida para seu objeto. Entretanto, valeu-se do conceito de meta-avaliação e da discussão sobre suas categorias para criar seu quadro de análise, apresentados nessa subseção.

2.3.1 O conceito de meta-avaliação

Meta-avaliar quer dizer avaliar a avaliação. O conceito foi proposto por Scriven em 1969 no contexto da avaliação educacional. Antes disso, havia alguma discussão sobre aspectos - especialmente os técnicos – de elementos da avaliação, como a qualidade dos instrumentos ou a escolha da abordagem metodológica, observada, por exemplo, nos padrões da APA (*APA technical standards for test development*, de 1954) ou o The *Burg´s Mental Measurement Yearbook,* de 1965. Contudo, pouco havia ainda sido publicado sobre delineamentos de meta-avaliação. Em 1974, Stufflebeam lança o documento Meta-avaliação (*Meta-evaluation*, pelo *Evaluation Center* da *Western Michigan Universtity*) no qual discute um arcabouço para a meta-avaliação inspirado nos critérios utilizados para analisar pesquisa científica e acrescidos da percepção do atendimento da necessidade dos *stakeholders* (STUFFLEBEAM, 1974). Como já mencionado na Introdução, em paralelo, três associações americanas (APA, AERA e NCME) formaram uma força tarefa que resultou na criação, em 1975, do *Joint Committee on Standards for Educational Evaluation* (JCSEE). Após a divulgação dos padrões para a avaliação educacional, em 1981, as três associações decidem que os mesmos precisam de manutenção e revisões periódicas e, eventualmente, de novos itens. Por essa razão, é criada uma organização com tal fim que, em 1989, passa a ser certificada, submetendo seus padrões ao *American National Standards Institute* (ANSI). Em 1994, há a divulgação dos padrões para avaliação de programa, aprovados pelo Instituto. São esses 30 padrões distribuídos em 04 categorias (precisão, viabilidade, propriedade e utilidade) que influenciam a criação de padrões por muitas associações e entidades de avaliação pelo mundo afora e que dão base para várias experiências de meta-avaliação. Ainda que para alguns autores, como Cook e Gruder (1978:5-7), o conceito devesse estar atrelado à avaliação de avaliações somativas, no presente trabalho adota-se a meta-avaliação de maneira mais ampla, incluindo as avaliações formativas, como proposto pelo JCSEE (1994) ou por Stufflebeam (1974, 2007).

No Brasil, a meta-avaliação tem aparecido na literatura mais recente[33] graças, principalmente, à expansão dos programas de avaliação educacional em larga escala e da demanda por avaliação de projetos e ações em diversas áreas (DAVOK, 2007; LETICHEVSKY *et alii*, 2005; 2007), especialmente na de Saúde (HARTZ, 2006). Ravela (RAVELA *et alii*, 2008) chama atenção para o fato de que uma avaliação sem qualidade, voltada apenas para a divulgação de resultados como mecanismo de prestação de contas, é uma falácia que pode fazer retroceder os sistemas de avaliação e impossibilitar uma discussão séria sobre *accountability*. De modo geral e especialmente para as avaliações *high stakes*, meta-avaliar é uma questão ética. Se algo está errado em uma avaliação, o erro não deve impactar, injustamente, o objeto da avaliação. Do mesmo modo, a meta-avaliação é considerada um reforço para o combate a (ou minimamente a identificação de) vieses e para a isenção político-partidária nas experiências avaliativas, ajudando a trazer-lhes transparência e rigor.

Meta-avaliar significa uma verificação sistemática de uma (ou mais) experiência (s) avaliativa(s) no sentido de determinar (e julgar) a qualidade de seus desenhos, processos e ou resultados (STUFFLEBEAM, 1974; STUFFLEBEAM, SHINKFIELD, 2007; COOKSY; CARACELLI, 2005; LEEUW, 2003; PENNA FIRME; LETICHEVSKY, 2002; REBOLLOSO *et alii*, 2002). Stufflebeam, ao propor seu primeiro modelo de meta-avaliação, já lembrava aos avaliadores que há sempre a possibilidade de algo estar errado na avaliação e que uma meta-avaliação pode identificar os problemas (como vieses, erros técnicos, má utilização) no delineamento, na implementação ou nos resultados encontrados. A questão nessa conceituação é, como em qualquer avaliação, identificar o que significa qualidade. Para associações de classe, instituições e experiências nas quais há padrões de qualidade para as avaliações e/ou avaliadores, o delineamento da meta-avaliação pretende observar o cumprimento de tais padrões (GIMENES, 2007; PENNA FIRME; LETICHEVSKY, 2002). Em instâncias nas quais os padrões não tiverem ainda sido propostos, a meta-avaliação, da mesma forma que a avaliação, precisa, como primeira etapa, estabelecê-los. Um exemplo recente: na área da Saúde, o texto de Hartz (2006) propõe padrões para a análise dos Estudos de Linha de Base do Projeto de Expansão e Consolidação da Saúde na Família exatamente pela ausência de normas nacionais, associativas ou governamentais, no Brasil. A ausência não é diferente na área da Educação.

De forma similar à avaliação, a meta-avaliação pode contribuir para o refinamento da experiência em curso (meta-avaliação formativa), para a discussão sobre avaliação no campo teórico, ou para

33 Em muitos casos, indevidamente, percebe-se que o pesquisador confunde os termos meta-avaliação e avaliação institucional.

ajudar novos desenhos avaliativos, como uma meta-avaliação somativa, para usar as expressões também propostas por Scriven nos meados dos 60[34]. A meta-avaliação se presta à responsabilização dos avaliadores e à prestação de contas e há uma expectativa de que colabore na tomada racional de decisões; pode ser usada também para auxiliar contratantes / demandantes de avaliação a decidirem-se por quem contratar. Leeuw lembra, com base no trabalho de Schwartz e Mayne (2003 *apud* LEEUW, 2003: s/p), que qualquer pessoa pode apresentar-se como avaliador, sem que o campo tenha um controle ou uma "carteira profissional" que ateste as características dessa pessoa (LEEUW, 2003).

Além disso, Cooksy e Caracelli (2005) argumentam que os produtos das meta-avaliações podem contribuir para a escolha de trabalhos que comporão avaliações feitas por meio da síntese de resultados de experiências avaliativas no campo (*evaluation syntheses*) ou para comparar os diferentes padrões de qualidade utilizados por avaliações ou conjuntos de avaliação. Dentre os vários usos da meta-avaliação, além obviamente da verificação da qualidade da avaliação primária, o que sobressai é o caráter de capacitação/formação daqueles sob avaliação (*learning organizations*), com impacto sobre os indivíduos e sobre suas organizações de trabalho. É possível também obter-se impacto da meta-avaliação sobre as organizações formadoras dos avaliadores (*capacity building*), exatamente por apontar os aspectos mais frágeis nas avaliações realizadas e, dessa maneira, ajudar a identificar aspectos do currículo que precisam de maior atenção (PENNA FIRME; LETICHEVSKY, 2002). No Brasil, onde o campo da avaliação é ainda incipiente, esse uso da meta-avaliação seria muito importante para consolidação da área.

No planejamento da meta-avaliação, é fundamental considerar as provisões para a implementação do delineamento proposto[35] e as variáveis externas que extrapolam o planejamento, mas que o afetam diretamente. Independente do delineamento proposto ou do uso a ser feito, a meta-avaliação é uma avaliação secundária de uma experiência avaliativa, denominada avaliação primária. O ciclo avaliativo poderia ser, desta forma, infinito: seria necessária uma avaliação terciária para investigar a secundária e assim por diante. Os teóricos da avaliação concordam que uma avaliação secundária para investigar uma primária é suficiente para a análise da experiência,

---

34 Stufflebeam (1974) utiliza as expressões proativa e reativa para designar a mesma idéia, no que é seguido por Davok (2007). Entretanto, especialmente a palavra reativa em Português pode assumir significados que falem de uma resistência ou uma não aceitação de algo. Por essa razão, para fins desse trabalho, optou-se por adotar a nomenclatura proposta por Scriven.

35 Ao relacionar a proposta de meta-avaliação com a abordagem do Marco Lógico (ORTEGÓN, PACHECO, PRIETO, 2005; BROSE, 2001), seria aconselhável incluir, na meta-avaliação, aquilo que no Marco Lógico é chamado "o pressuposto", i.e, as condições externas (dentre as quais as variáveis políticas) além das possibilidades de gerenciamento de um determinado projeto, mas para as quais o delineamento foi feito, que contribuem para sua realização e que podem ajudar a esclarecer quando algo não tiver acompanhado o desenho original.

propondo assim um fechamento para o ciclo avaliativo (STUFFLEBEAM, 1974). No ciclo das políticas públicas, a avaliação é uma das etapas. Desta forma, nada mais natural que avaliar as políticas de avaliação.

Para conceituar a meta-avaliação, é também interessante traçar algumas fronteiras, como feito anteriormente na seção que tratou da avaliação. A meta-avaliação pode ser plena, abordando todos os aspectos de uma avaliação primária, ou parcial, focando suas lentes sobre um ou outro de seus aspectos. Assim como qualquer avaliação, meta-avaliação não é pesquisa cientifica. Como qualquer avaliação, a meta-avaliação está a serviço de *stakeholders*[36] e de interesses específicos relativos a um determinado contexto, não havendo, portanto, uma obrigação, intrínseca ao conceito de meta-avaliação, com um ou outro posicionamento ideológico. Ao contrário, a meta-avaliação, assim como a avaliação primária, envolve valores diversos, de indivíduos, organizações e sociedades, que podem competir entre si. É aconselhável, entretanto, que os recursos empregados na meta-avaliação, da mesma forma que na avaliação, sejam revertidos – direta ou indiretamente – para a melhoria da experiência avaliada (um programa, uma ação, uma política - *evaluand* –), do indivíduo avaliado (*evaluee*)[37] e do bem comum. É fundamental que a meta-avaliação seja conduzida em atendimento a padrões e a princípios que garantam que as informações produzidas sejam válidas e precisas, em uma relação custo x benefício que a justifique.

Ao longo do tempo, os delineamentos de meta-avaliação variaram, assim como o próprio campo das ciências e em conseqüência de rápido avanço tecnológico[38], não havendo um formato certo ou errado (há um formato adequado). O próprio Stufflebeam inicia seu modelo de meta-avaliação focalizando o mérito do objeto (1974) e amplia o foco para mérito e valor nas abordagens mais recentes (STUFFLEBEAM; SHINKFIELD, 2007). De qualquer modo, ainda que tenham adjetivos e delineamentos diferentes, as estratégias usadas nos textos lidos apontam sempre para a busca da melhoria do processo avaliativo primário, de maneira que o mesmo possa contribuir para a qualidade do seu próprio objeto. A diferença entre o objeto da avaliação e o objeto da meta-avaliação é tratada na subseção a seguir.

36 No contexto da meta-avaliação, os avaliadores integrantes da equipe da avaliação primária estarão sempre nesse grupo de *stakeholders*.
37 Os termos *evaluee* e *evaluand* foram propostos por Scriven.
38 O avanço nas tecnologias de *hardware* computacional permitiu a proliferação de programas de leitura, tratamento e análise de dados, com impacto significativo para os delineamentos avaliativos.

2.3.2 A diferença de objeto entre a avaliação e a meta-avaliação

Nem sempre é conspícua a diferença entre o objeto da avaliação (o programa, a política, o indivíduo) e o objeto da meta-avaliação, ou seja, a avaliação primária. Estabelecer essa diferença é ainda mais interessante quando as categorias de análise incluem valor e mérito. A avaliação primária pode observar o valor e/ou o mérito do seu objeto. O Quadro 01 a seguir apresenta características de mérito e valor na proposta de Stufflebeam e Shinkfield (2007:10).

Mérito	Valor
Pode ser avaliado em qualquer objeto de interesse.	Apenas avaliado nos objetos com nível de qualidade já demonstrado e aceitável.
Avalia o valor intrínseco do objeto.	Avalia o valor extrínseco do objeto.
Avalia a qualidade, ou seja, o nível de excelência do objeto.	Avalia a qualidade do objeto e seu valor ou importância em um determinado contexto.
Pergunta: O objeto faz bem aquilo que deveria fazer?	Pergunta: o objeto tem alta qualidade e é algo que um grupo-alvo necessita?
Usa como referência padrões de qualidade aceitos para o tipo de objeto sob avaliação.	Usa como referência padrões de qualidade aceitos, bem como dados oriundos de levantamento de necessidades do grupo-alvo.
As conclusões classificam o objeto a partir dos padrões de qualidade e de comparação com outros objetos do mesmo tipo.	As conclusões observam o nível de qualidade aceitável do objeto e o classificam de acordo com a importância e o valor para um determinado grupo consumidor.
As avaliações de mérito podem ser feitas em termos de comparação do objeto com padrões ou com outros objetos.	As avaliações de valor podem ou não ser comparativas.

Quadro 1: Características de mérito e valor na avaliação primária, por Stufflebeam e Shinkfiel, 2007:10 (tradução deste autor).

Por exemplo, uma avaliação de um programa educacional em uma determinada prefeitura pode abordar o quanto a ação educacional é consistente no seu delineamento, foi ou está sendo implementada de maneira tecnicamente adequada, é viável em termos de custo x benefício e, principalmente, o quanto esse programa contribui para a melhoria (ou para o benefício) de uma determinada parcela da população que demanda aquele aspecto da educação. Em suma, a avaliação informa os *stakeholders* se o programa tem mérito e (ou) se tem valor. Os *stakeholders* aqui são aqueles do programa ou da ação: de modo direto, os formuladores e implementadores da política (bem como os contratantes da avaliação), a parcela da população interessada, e, de modo indireto, a sociedade em geral.

Já a meta-avaliação deve observar o valor e mérito da avaliação desse programa. Assim, ela vai responder questões concernentes ao delineamento da avaliação no atendimento à demanda dos *stakeholders;* se sua implementação é (ou foi) tecnicamente adequada de modo a resultar na "boa informação", se é viável e, principalmente, o quanto a avaliação foi útil (tanto em termos de

utilização real quanto do delineamento voltado para o atendimento dos *stakeholders*). Os *stakeholders* aqui são aqueles da avaliação: de modo direto, os formuladores e avaliadores da política (bem como os contratantes da meta-avaliação) e aqueles que sofreram impacto dos resultados da avaliação primária. De modo indireto, pode-se considerar o restante da comunidade interessada no programa educacional, a comunidade acadêmica envolvida com o tema da avaliação, e a sociedade em geral.

Na operacionalização de modelos de meta-avaliação da avaliação de programas e projetos, a diferença entre o objeto da avaliação (o programa) e o da meta-avaliação (a avaliação) nem sempre fica clara. Algumas das leituras feitas para a composição da fundamentação teórica da presente pesquisa mostram que os meta-avaliadores, em mais de uma situação, confundiram o seu papel e passaram a analisar valor e mérito do programa em si, e não da sua avaliação. Esse desacerto ocorre especialmente nos casos em que o meta-avaliador confunde avaliação com gestão e quer, ele mesmo, intervir *(*"*fix things*", segundo Stake, 2004:105). O quadro a seguir, adaptado no anterior, apresenta as características, consideradas na presente pesquisa, de mérito e valor para avaliações objeto das meta-avaliações, na busca por enfatizar que o *evaluand*, nesse caso, é a avaliação primária.

Mérito	**Valor**
Pode ser avaliado em qualquer experiência avaliativa, desde as isoladas até os sistemas de avaliação.	Apenas avaliado nas experiências avaliativas cujo mérito já tenha sido estabelecido.
Avalia o valor intrínseco da avaliação.	Avalia o valor extrínseco da avaliação.
Avalia a qualidade, ou seja, o nível de excelência da avaliação.	Avalia a utilidade (amplo sentido) da avaliação em um determinado contexto.
Pergunta: A avaliação atinge os objetivos para os quais foi delineada e implementada?	Pergunta: A avaliação atende às necessidades dos seus públicos-alvo?
Usa como referência padrões de qualidade aceitos como estado da arte em avaliação.	Usa também como referência os dados oriundos de levantamento de necessidades dos públicos-alvo.
As conclusões classificam a avaliação a partir dos padrões de qualidade e de comparação com outras experiências avaliativas.	As conclusões classificam a avaliação de acordo com a utilidade para seus públicos-alvo.
O mérito da avaliação pode ser estabelecido em termos comparativos, seja em relação aos padrões, ou ao estado da arte, ou ainda em relação a outras experiências.	O valor de uma avaliação pode ser estabelecido de modo comparativo ou não.

Quadro 2: Características de mérito e valor para meta-avaliação. Quadro proposto pelo autor a partir de Stufflebeam e Shinkfield (2007).

Mantidas as diferenças entre o objeto da avaliação e da meta-avaliação, o quadro muda um pouco quando a avaliação primária é uma política de avaliação. O ciclo de política pública pressupõe três etapas: formulação, implementação e avaliação. Quando a política sob análise é uma política de

avaliação, propõe-se aqui que a etapa de avaliação da política seja sua meta-avaliação, como pode ser percebido na ilustração a seguir.

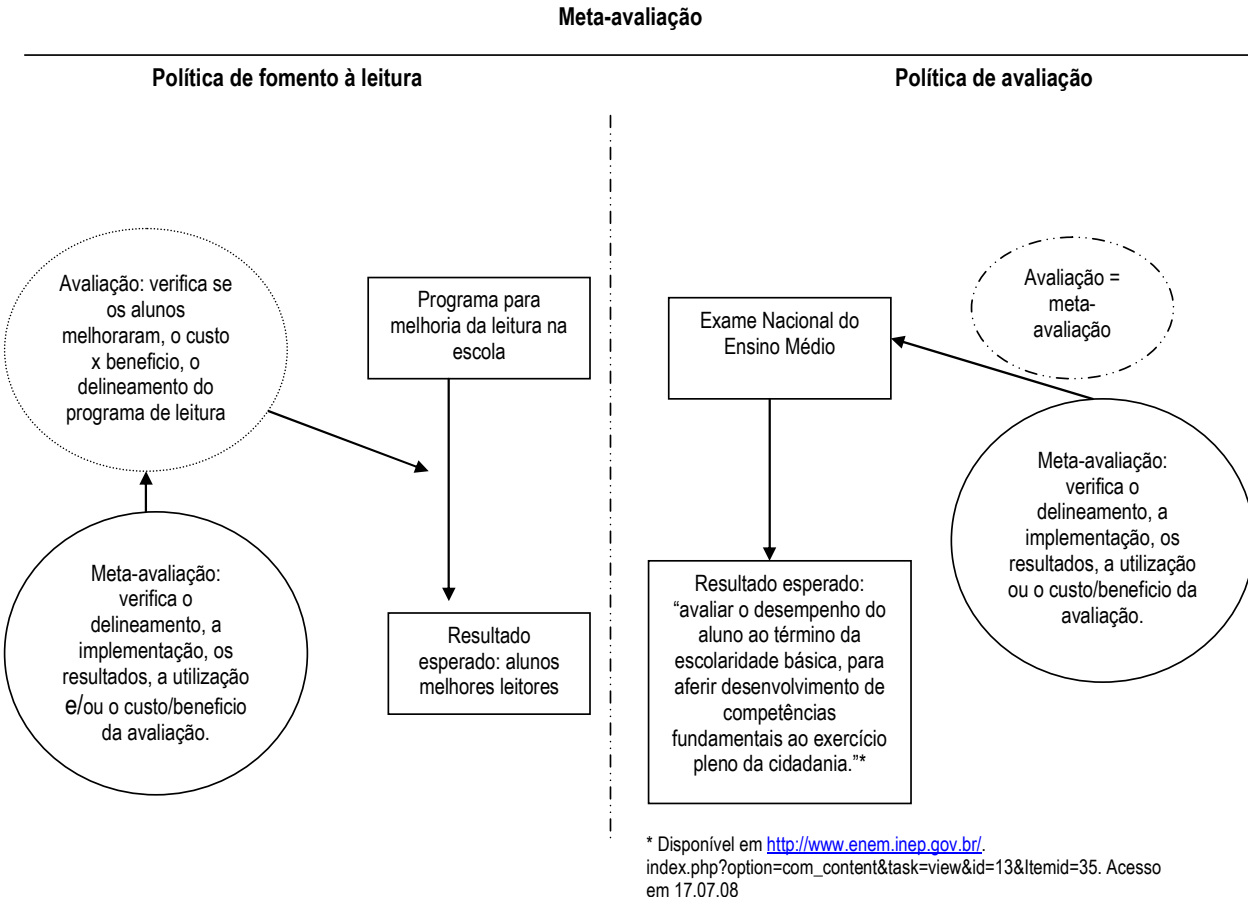


Ilustração 2: Exemplo para diferenciação da meta-avaliação quando a avaliação primária é uma política de avaliação e quando a avaliação primária focaliza outro tipo de política.

Usando o exemplo do ENEM, colocado na ilustração anterior, a sua meta-avaliação deve observar seu(s) mérito e/ou valor, com base nas definições dessa política. Assim, ela vai responder questões concernentes à avaliação da mesma forma que no caso anterior, mas os *stakeholders* são em maior grupo: de modo direto, os formuladores, implementadores e avaliadores da política (bem como os contratantes da meta-avaliação, se externa), os estudantes que fizeram as provas, as comunidades escolares de onde saíram esses estudantes e, em uma ampliação dos objetivos do ENEM, as universidades que utilizam seus dados no processo seletivo de entrada; de modo indireto, as secretarias de educação dos estados, o restante da comunidade educacional interessada na qualidade do Ensino Médio, em particular, e da Educação Básica, de modo geral; a comunidade acadêmica envolvida com Educação e com a avaliação e a sociedade em geral. Além disso, a meta-avaliação deverá considerar os aspectos de governança da política em seu delineamento, implementação e na própria meta-avaliação.

Conceituada a meta-avaliação e feita a distinção do seu objeto, a pergunta que se apresenta é: como delinear a meta-avaliação? É disso que trata a próxima subseção.

2.3.3 Questões cruciais para o delineamento de uma meta-avaliação

As mesmas perguntas que norteiam o delineamento de uma avaliação primária e os mesmos padrões de qualidade devem orientar o planejamento e a implementação de uma meta-avaliação. Contudo, quatro questões merecem um pouco mais de atenção: a escolha do meta-avaliador; a definição do momento do delineamento da meta-avaliação; as informações para o julgamento a ser feito; e o cuidado com as questões contratuais. As subseções a seguir discutem brevemente cada uma delas.

2.3.3.1 A escolha do meta-avaliador

Os elementos habitualmente discutidos para a escolha do avaliador cabem também para o meta-avaliador: as competências que o profissional deve dominar, as outras competências sobre as quais ele precisa ter algum conhecimento (ainda que não domine), o escopo de trabalho, os produtos e cronograma de trabalho, por exemplo. Um aspecto, no entanto, precisa de mais cuidado: o fato de o avaliador ser interno ou externo. Existe uma vasta discussão sobre essa escolha, que passa por resistência da equipe interna à equipe externa, dificuldade no tratamento das pressões, envolvimento com a “causa”, fontes principais ou secundárias de sustento para o avaliador externo, tempo que o avaliador externo tem contato com o objeto sendo avaliado, dentre muitas (YANG; SHEN, 2006; RAY, 2006; SHULHA; COUSINS, 1997).

Para Stufflebeam (1974), a avaliação deve ser conduzida por avaliadores internos e externos; geralmente (embora não sempre), os avaliadores internos conduzem a avaliação formativa para informar a tomada de decisões durante o processo, ao passo que avaliadores externos devem conduzir a avaliação somativa, que levará à responsabilização dos avaliadores. Essa posição leva em conta o refinamento da proposta e a responsabilização pelo trabalho, mas deixa de lado o repertório do avaliador ou da equipe da avaliação primária. Para a meta-avaliação, os contratantes deveriam contar com profissionais com espectro mais abrangente de competências e conhecimentos, externos, para que pudessem contribuir para o refinamento da avaliação primária. Parte-se da premissa de que o avaliador interno e a equipe da avaliação primária tenham usado da melhor forma o repertório que possuem para o desenvolvimento do trabalho. À equipe primária é importante um componente reflexivo durante toda a experiência que, aliás, deveria fazer parte de

qualquer condução de experiência. Mas, para a meta-avaliação poder discutir a avaliação primária, seria interessante um meta-avaliador (ou uma equipe de meta-avaliação) que pudesse questionar e contribuir para esse "repertório primário". Por essa razão, para a meta-avaliação, recomenda-se no presente trabalho que a equipe deva ser externa.

Sua escolha não é simples, no entanto. Como de resto em qualquer avaliação, a credibilidade é um fator de peso para que os resultados da meta-avaliação sejam considerados pelos avaliadores primários. É interessante perceber, por exemplo, um certo estranhamento entre os teóricos e acadêmicos da avaliação e os seus praticantes. Para os acadêmicos, os praticantes são os que concretizam experiências avaliativas (restritas à técnica) sem uma reflexão teórica, sendo, muitas vezes, rasos e acríticos. Para os praticantes, os acadêmicos vivem em um "mundo de livros", completamente distantes do "mundo real'. Se esses preconceitos não forem tratados minimamente durante a formação da equipe de meta-avaliação, é provável que os conflitos durante o desenvolvimento do trabalho levem a sua não utilização, em lugar de contribuir para o fortalecimento de ambas as equipes.

Além disso, o Brasil apresenta uma restrição em relação a essa escolha: sua comunidade de avaliação é incipiente, embora esteja crescendo e se capacitando. A escolha, portanto, é limitada, especialmente para a formação de equipes de meta-avaliação para as grandes políticas. Muito provavelmente, os luminares já estarão, de uma forma ou outra, envolvidos na equipe primária. Seu repertório, portanto, já terá sido colocado a serviço da avaliação primária e pode ser que, em alguns casos, seja difícil buscar alguém ou algum time de meta-avaliação cujo repertório favoreça as discussões e contribuições para a experiência primária. Em que pesem as diferenças culturais, uma sugestão é buscar meta-avaliadores fora do país, em locais onde as comunidades de avaliadores já estão mais consolidadas, como os EUA, a França ou, na América Latina, o Chile.

2.3.3.2 Definição do momento para o delineamento da meta-avaliação

Idealmente, a decisão da realização de uma meta-avaliação deve ocorrer concomitante à da avaliação, mesmo em caso de uma meta-avaliativa somativa, cujos esforços se concentrarão nas informações a serem fornecidas nos ou após os momentos finais da ação. Os delineamentos seriam então paralelos e os dados sobre a avaliação seriam produzidos e já organizados, durante sua formulação e implementação, de maneira a favorecer a meta-avaliação. Há duas vantagens nessa proposta: 1) a equipe da avaliação primária sabe, desde o início, que é avaliada e sobre quais critérios isso vai acontecer; assim não será surpreendida quando a meta-avaliação começar a buscar dados e a produzir resultados, o que reduz a resistência; e 2) o fato de os dados estarem

sendo organizados (no grau de desagregação ideal) implica otimização de recursos. O quadro a seguir ilustra esse desenvolvimento paralelo em caso de uma meta-avaliação formativa.

Etapas / Foco	Planejamento	Implementação	Finalização e Resultados	Conclusão
Foco da Avaliação	Elaboração do plano de avaliação integrado ao planejamento do seu objeto.	Levantamento, tratamento e análise dos dados.	Retorno das informações aos públicos-alvo. Contribuição na tomada de decisões e outros usos.	Finalização dos documentos sobre o programa. Reflexão sobre o plano implementado.
Foco da Meta-avaliação formativa	Elaboração do plano de meta-avaliação. Análise da consistência do plano de avaliação contrastado com o planejamento / discussão preliminar da viabilidade.	Análise da metodologia empregada e dos desvios do plano.	Análise do uso feito e da credibilidade do processo; conclusão sobre custo x efetividade.	Finalização dos documentos sobre a avaliação. Reflexão sobre a meta-avaliação implementada.

Quadro 3: Paralelismo nos momentos de avaliação e meta-avaliação com objetivo de otimização de recursos.

A diferença entre meta-avaliação formativa e somativa não está na definição do momento do seu planejamento ou da coleta de dados, mas nos tempos de devolução dos resultados e no grau de controle utilizado para a coleta. A desvantagem do paralelismo de implementação entre a avaliação e a meta-avaliação está na pressão exercida sobre a equipe da avaliação primária, que pode desviar seu foco do objeto sob avaliação para buscar as respostas à meta-avaliação. Ainda assim, é recomendável a decisão sobre a meta-avaliação quando do planejamento da avaliação.

Situações emergenciais, no entanto, justificam o delineamento de meta-avaliação não prevista no início do programa, na busca por validação de dados, identificação de problemas e de suas causas e, em extremo caso, até da anulação do processo avaliativo primário.

2.3.3.3 Informações para o julgamento a ser feito

O melhor cenário de trabalho para a meta-avaliação é aquele no qual os padrões e os critérios que indicam a qualidade da avaliação já foram definidos. Isso pode ocorrer, por exemplo, no texto de um edital para contratação de uma avaliação para um programa[39] ou, de maneira mais abrangente

39 Ver a já citada discussão de Hartz (2006) para o Programa Saúde na Família.

e geral, através de posicionamentos das instituições de classe ou da base legal. Não é o caso da avaliação educacional no Brasil.

Na ausência desse norte, cabe aos meta-avaliadores negociar os padrões de qualidade com os *stakeholders* (obviamente considerando o estado da arte da avaliação como fundamento), como primeiro passo da meta-avaliação. É preciso ter em mente que as diferentes visões do que seja qualidade da avaliação refletem as ideologias, os contextos e os campos de origem dos avaliadores. Portanto, é de crucial valor a determinação dos padrões para garantir a credibilidade do processo meta-avaliativo (COOKSY; CARACELLI, 2005).

Entretanto, habitualmente é esquecida a definição da linha de corte que separa a avaliação de qualidade daquela "sem qualidade". Por exemplo: de um modo geral, é fácil dizer que uma avaliação deve produzir dados precisos. É mais difícil determinar o grau de precisão que fará com que a avaliação seja julgada sem qualidade. Um outro exemplo: a avaliação será meta-avaliada por meio da verificação do cumprimento de um conjunto de 10 padrões. Há algum com maior peso que outro? Há algum que, isolado, defina a qualidade do conjunto? A análise será feita de modo analítico (um a um dos dez padrões) ou holístico (uma escala única que inclua todos os dez)? É tecnicamente possível definir uma escala única para padrões diferentes? Um bom exemplo da definição de uma escala pode ser visto em Penna Firme e Letichevsky (2002).

Diante dessas perguntas, vê-se que, tão importante quanto a definição dos padrões de qualidade, é a determinação dos critérios para julgá-los. Assim como os padrões, os critérios – ainda que atrelados a questões técnicas – precisam ser negociados com os *stakeholders* para evitar perda de credibilidade na meta-avaliação e conseqüente não utilização de seus resultados.

2.3.3.4 Questões contratuais

A literatura mostra a importância da meta-avaliação para a formação / capacitação das equipes de avaliação; muitas vezes, isso é referido como a "construção de uma massa crítica" para a avaliação. O último dos aspectos a serem discutidos em relação ao delineamento de uma meta-avaliação é a definição do escopo do trabalho a registrar em contrato em casos de meta-avaliação externa. É fundamental que sejam definidos, desde cedo, os produtos de trabalho da equipe de meta-avaliadores. A capacitação da equipe de avaliação primária é um desses produtos ou ele é resultado do trabalho da instituição contratante sobre os relatórios da avaliação? Essa questão é de suma importância porque há um limite a ser declarado sobre o que seja meta-avaliar, o que seja capacitar a equipe, e o que seja transferir tecnologia, por exemplo. A quantidade de horas a

utilizar na capacitação tem que ser considerada quando da formação do preço do trabalho. Da mesma maneira, a transferência de tecnologia implica normalmente um custo adicional à meta-avaliação.

Um segundo aspecto contratual a ser definido é a identificação da pessoa ou do grupo de pessoas da equipe primária que receberão produtos e serviços da equipe secundária. Queixas de que a equipe externa não informou a interna sobre os processos ou não entregou a base de dados ou que, de alguma forma, reteve informações são freqüentes em políticas de avaliação. Por essa razão, a propriedade da base de dados, a política de usos, e a política de divulgação atreladas a certa experiência avaliativa (seja primária ou secundária) precisam estar definidas contratualmente, para evitar problemas posteriores.

Da perspectiva desse trabalho, são essas as quatro questões (escolha do meta-avaliador, momento de delineamento, base do julgamento a ser feito, e questões contratuais) que merecem grande cuidado no delineamento da meta-avaliação, obviamente não se tendo descartado as demais (discutidas na seção Avaliação). Ficam no ar ainda algumas preocupações: que de maneira a meta-avaliação olha para a avaliação primária? Como abordar uma experiência complexa como a avaliativa sem retalhá-la? Como diferenciar as preocupações com um ou outro aspecto avaliativo (Os instrumentos foram bem construídos? Houve tratamento dos dados brutos?, por exemplo) de preocupações mais holísticas como: O delineamento da avaliação levou em consideração as demandas dos *stakeholders*? Ou, a avaliação é justificável em termos de custo? Ainda em 1974, Stufflebeam (1974:5) defendia que, ao indicar uma metodologia para a meta-avaliação, era importante ter em mente um conjunto apropriado de critérios, necessários para identificar atributos básicos e suficientes nos relatórios de avaliação e nos seus delineamentos. Em 1999, propôs um *checklist* para a meta-avaliação de avaliações de programa, no qual identificou as subcategorias mais importantes e os *score*s para cada uma das quatro categorias do JCSEE, de modo a permitir identificação das avaliações que minimamente tivessem conseguido atingir tais padrões de qualidade. No recorte desta pesquisa, entretanto, o foco é dado sobre a categoria Utilidade, complementada pela categoria Uso. É sobre elas, representando no seu conjunto as contribuições das políticas da avaliação, que trata a seção a seguir.

2.4 Contribuições da avaliação

Para o presente trabalho, as discussões sobre a utilidade e os usos da avaliação (para além dos resultados), genericamente entendidas como suas contribuições, são a base do quadro de análise, que se vale da categoria Utilidade (adaptada dos padrões do JCSEE) e propõe uma nova categoria, Uso, a partir da discussão entre Patton e Weiss aprofundada a seguir. Do posicionamento de Weiss sobre os usos da avaliação no contexto das políticas públicas, também discutido a seguir, foi derivada a hipótese de pesquisa. Focalizando as contribuições da política de avaliação sob a presente investigação, pressupõe-se que os aspectos técnicos (Precisão), de viabilidade (categoria Viabilidade) e aqueles da categoria Propriedade, apresentados na Subseção 2.3.1, tenham atingido um nível de qualidade minimamente suficiente.

Essa seção é iniciada com um panorama dos estudos sobre usos da avaliação, no qual se percebe uma mudança no foco de investigação ao longo do tempo. Antes voltada para os fatores preditores do uso, as pesquisas sobre uso passam a extrapolar os usos dos resultados e a envolver usos de processo. Além disso, tais estudos voltam-se também para outros tipos de interação entre a avaliação e a realidade avaliada.

A partir desse panorama, essa seção traz outras duas subseções, correspondentes às categorias de análise do presente estudo: Utilidade e Uso. O detalhamento da categoria Utilidade do JCSEE teve por base o *checklist* de verificação da qualidade de avaliação de programa de Stufflebeam (1999). Um esforço foi feito no sentido de aproximá-la da avaliação de política de avaliação (incluindo uma análise da percepção de *stakeholders*, não proposta originalmente). Em seguida, a partir da discussão sobre tipos de uso, foi apresentada a categoria Uso. Do conjunto das discussões sobre Utilidade e Uso, foi possível construir o modelo de análise utilizado para o estudo em tela, detalhado quando da apresentação do capítulo 3. Metodologia.

2.4.1 Panorama de estudos sobre os usos da avaliação

A preocupação com o uso da avaliação, de início atrelada apenas aos seus resultados, é antiga e, muitas vezes, justificada pela observação da sua falta ou de maus usos. Há um número

considerável de autores[40] que se dedicam a esse tema, como Patton, House, Chelimsky, Shulha, Cousins, Weiss, Leviton, Henry, Lawrenz, Gullickson e Toal, alguns desde a década de 60.

De acordo com Lawrenz, Gullickson e Toal (2007), os estudos sobre o uso da avaliação concentravam-se, inicialmente, em torno dos modos pelos quais os decisores utilizavam os resultados da avaliação, dos fatores que impactavam o uso e das conseqüências desse uso a longo prazo. Segundo Shulha e Cousins (1997), em um apanhado histórico sobre os estudos da avaliação, por volta de 1986 havia uma noção geral de que uso é um fenômeno multidimensional melhor descrito a partir das categorias instrumental, conceitual e simbólica. Para incentivar e compreender o uso da avaliação, produziam-se listas de preditores, como, por exemplo: relevância, credibilidade, envolvimento do usuário, eficácia na comunicação, potencial de processamento de informação, demanda do cliente por informação, grau antecipado de mudança no programa, percepção da avaliação como ferramenta de gestão, qualidade da implementação da avaliação, e características contextuais do processo decisório (a categoria Utilidade do JCSEE está, em certa medida, inserida nesse contexto de estudos sobre elementos preditores de uso). Inicialmente uma descrição de fatores, os estudos evoluíram para incluir levantamento dos pesos relativos de cada fator no seu aspecto preditor de uso.

Muitos anos depois, em 2006, autores como Bamberger, Ruth e Madry (2006) continuam discutindo os fatores que impactam o uso. Para eles, o não atendimento ao *stakeholder* no tempo certo, a falta de flexibilidade e de efetividade dos delineamentos avaliativos em relação às necessidades dos *stakeholders*-chave, perguntas avaliativas erradas e achados irrelevantes, metodologia fraca, alta demanda por parte dos avaliadores em termos de recursos (financeiros, humanos, etc.) e a falta de *expertise* local para a condução, revisão e utilização da avaliação são os principais fatores para a baixa ou para a ausência de utilização da avaliação.

Não traz espanto, portanto, que, para autores como Leviton (2003), o campo dos estudos sobre os usos da avaliação tenha mudado pouco desde a proposição de Patton, originalmente posta em 1976-78. Patton tem defendido desde então que é o uso que justifica uma ação avaliativa e, nesse sentido, é necessário que seu delineamento assegure usos determinados para usuários definidos (*intended uses by intended users*). O autor considera o avaliador responsável pela concretização do uso e advoga que o envolvimento e a co-participação dos *stakeholders* são peças fundamentais para que o uso ocorra (PATTON, 1997).

---

40 Ver as referências de Patton (1988; 1997), Shulha; Cousins,(1997), Weiss (1988), Leviton (2003), Henry (2003), e Lawrenz; Gullickson; Toal (2007).

Em uma clássica palestra feita em 1988 e transcrita para o *American Journal of Evaluation*, Patton (1988) propõe que os avaliadores também sejam avaliados e imputados caso suas avaliações não resultem em usos. Muitos avaliadores, como ele próprio, dedicaram horas de trabalho a, além de propor metodologias para responder as perguntas avaliativas de seus clientes, identificar previamente os tipos de uso que poderiam ser feitos pelos *stakeholders* e usuários e a delinear as avaliações de modo a facilitar tais usos. A avaliação passava a ser vista como um diálogo contínuo e propunha o compartilhamento, entre avaliadores e avaliados/contratantes, da responsabilidade pela coleta, processamento e consumo da informação (SHULHA; COUSINS,1997). É importante lembrar que, para grande parte desses autores, o lugar onde se colocam é aquele da avaliação de programas, em pequena ou larga escala.

Weiss (1998) mantém uma posição diferente da de Leviton e sua discussão com Patton sobre a responsabilidade do avaliador para a concretização do uso da avaliação contribuiu muito para o enriquecimento do campo de estudos[41]. Para ela, originária dos estudos sobre políticas públicas e cujo interesse é avaliação de políticas, houve evolução dos estudos sobre os usos da avaliação. Essa evolução se deu porque foram inseridos novos construtos e perspectivas, vez que os pesquisadores e avaliadores, ao compreender a complexidade do fenômeno avaliativo, expandiram seu foco das características da avaliação, dos usuários potenciais e das estratégias de comunicação para observar também a interação entre tais características, o avaliador, o contexto, e os métodos de comunicação envolvidos. Weiss (1998:27), citando Breslau, diz que a pesquisa já mostrou que o uso da avaliação inclui as categorias de dados, o desenho da avaliação e da análise como elementos de uso. Os antes chamados efeitos colaterais da avaliação e mesmo alguns dos maus usos deixam assim de ser entendidos como desvios ou aspectos não desejados e passam a ser percebidos como usos diversos daqueles originalmente planejados (inclusive tornando-se objetos de novos estudos). Nessa perspectiva, boa parte dos aspectos de tais interações não pode ser determinada pelo avaliador, que precisa delinear uma avaliação voltada para o uso, mas que não é o único responsável caso sua concretização não se dê como programada.

Patton (1988; 1997) discorda integralmente dessa posição, por acreditar que é possível para o avaliador definir perguntas avaliativas em conjunto com *stakeholders* e usuários e, nesse delineamento, obter o compromisso pelo uso programado dos resultados. O avaliador deve ser,

41 A discussão entre Patton e Weiss pode ser acompanhada em uma série de artigos publicados *no American Journal of evaluation* e no *Evaluation*. Os dois artigos mais citados são o de Patton, *The evaluator´s responsibility for utilization* (1988) e os de Weiss, *Is anybody there? Does anybody care?* (1987) e, onze anos mais tarde, *Have we Have we learned anything new about the use of evaluation*? (1998)

portanto, imputável pelo uso instrumental, aquele que leva imediatamente e diretamente à melhoria da qualidade do objeto sob avaliação.

Tanto Patton quanto Weiss têm razão. Enquanto o primeiro tem a avaliação de programa (independente da escala) como objeto de interesse, estudo e prática, a segunda se preocupa com a avaliação de políticas. As características do objeto são, nesse cenário, definidoras dos graus de imputabilidade do avaliador em relação à concretização do uso. Concorda-se com Patton quanto à importância do uso instrumental como a razão de ser da avaliação, especialmente em um contexto de reforma do Estado, onde se busca a otimização das ações, como visto na primeira seção do Marco Teórico. Por outro lado, Weiss é sábia ao reconhecer que nem sempre as condições de contexto político são estáveis o suficiente para garantir esse uso, mas que há outros importantes efeitos da avaliação, observáveis no longo prazo e nem sempre atrelados ao uso instrumental.

Em que pese a imputabilidade (ou não) dos avaliadores pelo uso direto da avaliação, seus estudiosos ampliaram o leque de interesses: em lugar de restringir o uso à fase dos resultados (ou achados), passam a observar, por exemplo, outras etapas de utilização, como o planejamento, a implementação e finalização de uma experiência avaliativa. Segundo Weiss (1998), são vários os elementos da avaliação usados: os achados (resultados), as recomendações (se e quando existentes), as idéias e generalizações, o processo, a discussão. Shulha e Cousins (1997) relatam que o reconhecimento e a aceitação de uso de processo abriram o campo para novas abordagens avaliativas e contribuíram para a elaboração de novas questões de avaliação. Os avaliadores que trilharam esse caminho focalizaram, em grande parte, a forma como o delineamento e as práticas da avaliação impactaram a aprendizagem individual ou organizacional, como relatam Forss, Renien e Carlsson (2002).

Weiss (1998) argumenta, por exemplo, que o simples fato de se estar sob avaliação pode impactar o indivíduo, a organização ou o programa/política positiva ou negativamente. Positivamente porque denota uma preocupação com melhoria e com transparência, com prestação de contas, e, de alguma maneira, essa imagem acaba por contribuir para a legitimação do programa. Essa percepção é compartilhada por muitos avaliadores em contato com *stakeholders* governamentais durante a negociação da política de avaliação. De alguma maneira, ao falarmos de sistemas de avaliação educacional, vê-se esse uso político: o Estado y ou x implementou um sistema de avaliação para mostrar que está acompanhando suas ações e para prestar contas à sociedade, independente do uso ou do não-uso dos resultados da avaliação para a tomada de decisões ou formulação de novas políticas (FERRER, 1997; WEISS, 1998; RAVELA *et alii*, 2008).

Negativamente porque, devido a uma cultura ainda existente, há uma percepção de que se algo está sendo avaliado é porque tem algum problema.

No cenário da avaliação de políticas, fatores para a baixa utilização não necessariamente são devidos ao delineamento da avaliação. Eles podem estar relacionados às crenças conflitantes na equipe do programa sob avaliação, ao posicionamento ideológico ou partidário das lideranças, aos interesses divergentes entre unidades do programa, à incapacidade na obtenção de consenso quanto ao escopo a ser avaliado, à rigidez das regras organizacionais, às mudanças no ambiente externo (como os cortes orçamentários), dentre outras questões fundamentais (WEISS, 1998). O argumento que Patton (1988) utiliza para discordar de Weiss é que o avaliador já conhece esses problemas e que, ao delinear a avaliação, deve já endereçá-los de maneira a evitar que ocorram. São, portanto, duas correntes: a primeira, de Patton, pressupõe que o avaliador deva ser imputável pelo uso instrumental da avaliação, independente das condições contextuais; a segunda, na qual Weiss é expoente, advoga que há aspectos contextuais que interferem no uso da avaliação, que estes estão fora do raio de influência do avaliador, e que outros usos, não previstos e não instrumentais, devem ser observados na análise das avaliações. As duas correntes apresentadas aqui contribuem para a compreensão do que ocorre para que as experiências avaliativas sejam úteis.

A hipótese apresentada na presente pesquisa

> em políticas de avaliação educacional em larga escala, os resultados são elementos pouco utilizados e é o acontecimento da avaliação que afeta as instituições em nível micro (escolas)

é derivada da posição de Weiss e demais autores que ampliaram o entendimento do uso para além da fase de utilização dos resultados e para além do uso instrumental, discutida nos parágrafos anteriores. Essa escolha foi corroborada pelo trabalho de mestrado do autor (DANTAS, 2005), no qual foram levantados desvios de implementação da política de Avaliação da Aprendizagem, foco também do estudo em tela. Em muitos dos casos observados durante o mestrado, esses desvios poderiam ser relacionados a usos inesperados – nem sempre instrumentais e muitas vezes não relacionados aos resultados da avaliação - não originalmente previstos na formulação da política. Esses usos e utilidade são aqui considerados contribuições da avaliação.

Mais recentemente (década de 00), observa-se a inclusão da influência (em indivíduos, programas e comunidades) como categoria de pesquisa, expandindo a noção de uso no tempo e no contexto

(HENRY; MARK, 2003; LAWRENZ; GULLICKSON; TOAL, 2007:276). Nessa corrente, Kirkhart (2000 *apud* LAWRENZ; GULLICKSON; TOAL, 2007:276) propõe três dimensões de influência: fonte, intenção e tempo. Tanto o processo quanto os produtos da avaliação (fontes) informam e afetam uma pessoa de maneira intencional ou não (intenção), durante, imediatamente após ou muito tempo após a finalização da avaliação (tempo). Dessa maneira, estudos sobre a influência da avaliação podem abranger um leque maior de conseqüências da experiência avaliativa.

Para Henry e Mark (2003), os estudiosos devem olhar para "além do uso" e mais explicitamente considerar a influência como o elo de ligação entre a avaliação e seu impacto. Para esses autores, ao propor uma teoria das mudanças, o programa de avaliação em si é uma intervenção social. Henry e Mark fazem parte do grupo de pesquisadores sobre avaliação que, no início de 00, associavam a finalidade maior da avaliação à melhoria social (*social betterment*) e, nesse contexto, a influência, mais abrangente que o uso da avaliação, torna-se a ferramenta pela qual se atinge tal melhoramento (HENRY; MARK, 2003: LAWRENZ; GULLICKSON; TOAL, 2007).

Transpondo a influência dos programas em si para a esfera de políticas públicas, um exemplo é o SAEB. Essa política de avaliação em larga escala, no Brasil, foi amplamente usada como modelo ou, minimamente, como incentivo à adoção, por parte dos Estados, de programas de avaliação, especialmente na segunda metade da década de 90, quando se consolidou. Para aqueles estados como Pernambuco, São Paulo ou Ceará, com políticas avaliativas anteriores ao SAEB, a avaliação federal influenciou o delineamento adotado, ao menos, na freqüência de aplicação de provas (como o SAEB fazia aplicação em anos impares, os Estados começaram a se mobilizar nos anos pares). Para outros Estados, como a Bahia em 2000, o SAEB foi uma referência. Aos poucos, os representantes dos Estados brasileiros começaram a discutir as políticas de avaliação, em uma formação – ainda que bastante lenta – da cultura de avaliação no Brasil (BONAMINO *et alii*, 2004).

Novamente, apresenta-se aqui a crítica de Patton e daqueles que adotam seu posicionamento: a influência é um conceito abstrato e aberto e não pode ser diretamente atrelado à tomada de decisões que levem ao melhoramento de um *evaluand*. Por essa razão, em termos de avaliação, essa categoria não é "vendável". Mais uma vez, cabe diferenciar avaliação de programa de avaliação de política. No primeiro caso, os *stakeholders* principais são, comumente, associados a instituições privadas (mesmo que sem fins lucrativos) interessadas em fazer seus investimentos na área social valerem o máximo no menor tempo possível. No caso das políticas públicas, nem

sempre isso é possível. Os governos, até aqueles mais "presos" aos períodos de seus mandatos ou ao discurso de otimização das reformas, tendem a pensar no bem comum a longo prazo como finalidade maior. Especialmente quanto às políticas, os efeitos da influência da avaliação, mesmo que não as justifiquem, devem ser considerados. Infelizmente, dadas as limitações do presente estudo, a influência não foi observada como categoria de análise da política de avaliação de aprendizagem.

No Brasil, a literatura encontrada não refere a estudos específicos sobre o uso ou a utilidade da avaliação educacional por si, embora uma série de autores, como Mere Abramowicz, Sandra Zákia Souza, Romualdo Portela[42], Robert Verhine, José Dias Sobrinho, Alicia Bonamino, Creso Franco, Ana Carolina Letichevsky, Tereza Penna Firme, Cláudio Moura Castro e Simon Schwartzman, em textos relacionados nas referências do presente trabalho, de uma maneira ou de outra, discutam como as avaliações impactam ou deixam de impactar um determinado segmento educacional, em geral referindo-se a avaliações de grande escala, como o SAEB, o Provão, o ENADE, o modelo CAPES ou modelos implementados pelos diversos Estados da Federação, ou a sistemas de avaliação, como o SINAES. Em muitos desses trabalhos, a discussão se faz sobre a finalidade da avaliação e, tangencialmente, refere-se aos usos. Por vezes, percebe-se que uso e utilidade ou uso e finalidade são usados sem distinção (o próprio JCSEE usa utilidade para nomear uma categoria que trata de elementos preditores de uso). Finalidade, utilidade e uso são termos diversos e devem ser tratados diferentemente[43]. Para facilitar o entendimento das políticas de avaliação, é necessário que se faça uma distinção entre eles.

A. Finalidade

A finalidade de uma política de avaliação é sua contribuição esperada, normalmente traduzida pelo objetivo geral no programa ou projeto que a implementa. Reforça-se aqui o entendimento de que um programa ou um projeto são políticas postas em prática. A ampla abrangência dos objetivos já foi tratada na subseção que relata a centralidade da avaliação. No caso das políticas de avaliação educacional, o objetivo geral vem freqüentemente associado a uma declaração, muito além do escopo da avaliação, de contribuição para a qualidade da educação. Esse objetivo geral-finalidade é, em verdade, pensado como um grande objetivo comum às diversas políticas de um determinado programa de governo que, em articulação e desde que bem delineadas, impactarão a

42 Romualdo Portela Oliveira e Sandra Zakia Sousa têm, desde 2001, desenvolvido trabalho de pesquisa sobre as políticas de avaliação no Brasil e o seu uso, de acordo com informações no Lattes de Portela. Entretanto, até o início de 2009, não havia sido divulgado qualquer artigo ou publicação desses pesquisadores sobre o tema.

realidade se implementadas em conjunto. Por exemplo, de acordo com Sousa (2003: 180), o SAEB tem "como finalidade reverter o quadro de baixa qualidade e produtividade do ensino, caracterizado, essencialmente, pelos índices de repetência e evasão escolar." É claro que a finalidade definida para o SAEB é muitas vezes superior às possibilidades de impacto isolado de uma política de avaliação, mas o conjunto das políticas educacionais de um determinado governo ou do Estado pode, se e quando articulado, contribuir para seu atingimento.

Um segundo exemplo: a Lei federal 10.861, de 14 de abril de 2004, ao instituir o SINAES (Art. 1º, § 1º), determina que suas finalidades são:

> a melhoria da qualidade da educação superior, a orientação da expansão da sua oferta, o aumento permanente da sua eficácia institucional e efetividade acadêmica e social e, especialmente, a promoção do aprofundamento dos compromissos e responsabilidades sociais das instituições de educação superior, por meio da valorização de sua missão pública, da promoção dos valores democráticos, do respeito à diferença e à diversidade, da afirmação da autonomia e da identidade institucional. (BRASIL, 2004)

Como visto antes sobre o SAEB, o SINAES isoladamente não tem como impactar a melhoria da qualidade da educação, como de resto qualquer avaliação. Para Lipsky (1980), os objetivos gerais das políticas públicas são abrangentes demais, o que os torna distantes da concretização. Essa amplitude dificulta enormemente a definição do quê avaliar. Considerando a avaliação de programas, Patton (1997) argumenta que as finalidades podem ser agrupadas em três categorias: as avaliações voltadas para o julgamento do *evaluand* (como as avaliações somativas de Scriven), aquelas voltadas para o melhoramento do *evaluand* (como as formativas) e os delineamentos para produção de conhecimento. Essas três finalidades maiores vão determinar os usos em cada contexto e os usos da avaliação deverão ser voltados para o atingimento dessas finalidades. Para as políticas de avaliação no Brasil, pode-se pensar nas três categorias de Patton, tomando como *evaluand* o próprio sistema educacional, seja na esfera federal, seja nas demais esferas.

Para dificultar ainda mais o atrelamento da qualidade da avaliação à(s) sua(s) finalidade(s), em alguns delineamentos, há uma confusão entre gestor e avaliador e, noutros, espera-se que o avaliador seja também interventor. Essa confusão é problemática, especialmente nos casos de avaliação em longos períodos de tempo. O avaliador gestor sairá do foco avaliativo para se envolver na tomada de decisões e o avaliador interventor perderá a perspectiva ao se envolver

---

43 No presente trabalho, optou-se por manter a nomenclatura do JCSEE para sua categoria Utilidade, mas entende-se que ela não se refere à utilidade percebida por *stakeholders* como elemento de análise da qualidade de experiências avaliativas.

diretamente com a ação sob avaliação. Em uma segunda rodada avaliativa, esse avaliador avaliará a si mesmo.

Os objetivos específicos retratam melhor a contribuição direta de uma determinada experiência avaliativa. Talvez o valor da experiência avaliativa deva ser buscado muito mais pelos seus usos e utilidade atrelados aos objetivos específicos que no cumprimento da finalidade maior do programa educacional, dada sua amplitude. Para Ginsburg e Rhett (2003), o valor de uma nova avaliação é determinado pelo adicional de informação que ela provê ao campo, já que há hoje um corpo de evidências científicas acumulado. As informações podem confirmar achados anteriores, oferecer novos entendimentos sobre os programas e políticas ou ainda colocar em cheque premissas e pressupostos sobre intervenções em particular. Os delineamentos de avaliações de políticas não garantem os usos, mas devem ser tais que aumentem a probabilidade de que eles ocorram. Para esses autores, a avaliação de programa educacional deve ser delineada de modo a contribuir com evidências relevantes que aumentem a probabilidade de as decisões tomadas colaborarem para a melhoria do desempenho do programa.

Finalidade e uso são, portanto, conceitos diferentes. No modelo de análise da presente investigação, a finalidade foi contemplada como o indicador “atingimento dos objetivos da política”, dentro da categoria Uso. De alguma maneira, espera-se que, caso a avaliação seja usada, os objetivos sejam atingidos.

Vale lembrar que, no presente documento, os termos uso e utilidade estão em letra minúscula quando se referem aos conceitos de maneira genérica e que, quando escritos com maiúscula (Utilidade e Uso), referem-se às categorias de análise da pesquisa. Antes de passar para a conceituação do termo uso, é importante que seja conceituado o termo utilidade, o que será feito a seguir.

B. Utilidade

Os parágrafos anteriores delimitaram o termo finalidade. O segundo termo, utilidade, é atrelado à percepção de *stakeholders* e usuários. Independe da finalidade da avaliação e relaciona-se com a demanda de informações (e, por vezes, não apenas de informações, mas infra-estrutura, recursos, etc.,) que cada um deles tem. Normalmente, a política de avaliação atende a um recorte dessa demanda (já que há interesses diversos envolvidos em um contexto de limitações de tempo, custo e técnica, como discutido anteriormente). Os formuladores da política de avaliação educacional são *stakeholders* principais com demandas conflitantes, por vezes, com outros *stakeholders* e

usuários, como, por exemplo, a comunidade docente nas escolas avaliadas ou o sindicato dos professores. Concorda-se com Ferrer (1997:2) quando diz: *"las expectativas que uno deposita en la actividad evaluadora, el papel que le otorga y la perspectiva que adopta para llevar a cabo sus análisis condicionan el juicio de utilidad que finalmente emite"*.

É a etapa de planejamento político-conceitual, no delineamento da experiência avaliativa, o ponto de partida para a negociação entre os *stakeholders.* Pretende-se que as expectativas sejam acomodadas pelo que a realidade mostra que é possível realizar, mas raramente se atinge um consenso. Para o grupo de *stakeholders* que ficou fora do atendimento, a avaliação, apesar de ter tido muitos usos, não foi útil. Para dar maior complexidade a essa discussão, há ainda a questão do custo: como posto por Stufflebeam (1974), os achados da avaliação devem valer mais para seus públicos que o custo de obtenção da informação. A questão enfrentada pelas políticas de avaliação é instigante: como afetam um grande número de *stakeholders*, para alguns o custo é justificado e, para outros, não, em um mesmo contexto e sob o mesmo delineamento avaliativo. De acordo com Chelimsky (1983), "o conceito de utilidade depende da perspectiva e dos valores do observador. Aquilo que para um é útil pode ser desperdício para outro" (1983:155 *apud* PATTON 1997:64, tradução deste autor).

A investigação sobre a utilidade é também interessante nos casos em que a percepção está atrelada fundamentalmente a posições ideológicas do indivíduo (ou seu grupo), o que o faz discordar da abordagem ou dos objetivos da política de avaliação. Nesse caso, a expectativa que se tem da avaliação é de que seja uma ferramenta a serviço de uma ideologia outra que não a sua, o que a torna não útil e, por vezes, deletéria. É o caso, por exemplo, da discussão avaliação x regulação apresentada no documento de proposição do SINAES. Alguns membros do grupo propositor declaram inadequada a abordagem anterior ao SINAES por prestar-se a "informar o mercado", vez que os *rankings* do Provão foram utilizados para propaganda dos cursos privados com melhores notas.

Além disso, em muitos sistemas educacionais, pressupõe-se que a mera realização da avaliação, sem articulação governamental e/ou uma posterior intervenção no sentido de endereçar os pontos fracos por ela apontados, resulta em melhoria da qualidade educacional (VIANNA, 1998; RAVELA *et alii*, 2008). Exige-se mais da avaliação do que é capaz de dar, por um lado, e confunde-se avaliação com gestão, por outro. Nesses momentos, a percepção será sempre de uma avaliação pouco útil.

Outro elemento que pode afetar a percepção de utilidade é o “medo da avaliação” que, em muitos casos, tem justificado a resistência à política e o seu não uso, especialmente em casos de avaliação *high stakes*. Esse é um fenômeno a investigar nas políticas de avaliação no Brasil. Raramente elas são delineadas de modo a impactar o indivíduo e, mesmo quando se prestam à regulação, são planejadas com uma série de etapas para que o *evaluand* se ajuste, com baixo impacto. O medo, portanto, não seria justificado pelo caráter *high stakes*, mas é referido e afeta a percepção de utilidade.

Para Rebolloso *et alii* (2002), a utilidade envolve quatro componentes analisáveis: a relevância dos dados em relação às necessidades de informação, o tempo de entrega dos resultados da avaliação, a apresentação dos relatórios (entendidos com a organização lógica da informação de modo que seja coerente e facilmente compreendida), e o impacto real desse relato (em termos de uso demonstrável ou de sua influência sobre a legislação ou sobre o processo decisório). Para os autores, as medidas de utilidade visam ao fortalecimento da execução de negociações específicas com o cliente e o monitoramento da satisfação e do uso. Os componentes são próximos de alguns dos itens verificadores da categoria Utilidade do JCSEE e, embora a proposta de Rebolloso *et alii* observe certa mistura dos conceitos de uso e utilidade, é interessante observar que esses autores trabalham com o “monitoramento da satisfação”, em clara associação da utilidade à percepção de usuários e *stakeholders*.

De maneira geral, quando a categoria Utilidade do JCSEE aborda a identificação dos *stakeholders*, o levantamento de suas demandas, as negociações e a priorização de alguns, trata de elementos que afetam a percepção da utilidade. Entretanto, a categoria não propõe um levantamento dessa percepção. Advoga-se aqui que tal percepção, por um lado, é resultante de um uso feito (ou resultante do atendimento das expectativas dos *stakeholders*), mas, por outro lado, também é preditora de uso: se há uma impressão geral de que haverá o atendimento de expectativas, então o *stakeholder* buscará ficar mais perto da experiência avaliativa, o que leva ao uso. Por essa razão, para fins do presente estudo, são utilizados dados da percepção de utilidade dos *stakeholders/usuários* principais, os gestores das unidades escolares. Esse é o U8 da categoria Utilidade. Conceituados finalidade e utilidade, falta apenas a definição do termo uso.

C. Uso

A terceira conceituação relaciona-se ao termo uso. Usos são as ações desenvolvidas pelos *stakeholders* e usuários (ou atitudes adotadas por eles) a partir de elementos da avaliação e, muitas vezes, ultrapassam ou se desviam da finalidade da avaliação; podem ocorrer em nível individual e

coletivo, desde o planejamento da avaliação até sua finalização. Como já visto, nos primeiros tempos de estudos sobre a avaliação, eram os achados que determinavam os usos. Após o posicionamento de Patton em 1976/1978, busca-se antes a demanda dos usuários para que o delineamento avaliativo leve a achados que venham a ser utilizados (PATTON, 1997).

O avaliador, ao delinear a política de avaliação, o faz para facilitar e incentivar determinados usos; a política sob implementação pode sofrer variações que levam a outros processos e usos (DANTAS, 2005). Weiss (1999), discutindo usos, apresenta pesquisas conduzidas, por exemplo, por Hocking (1988 *apud* WEISS, 1999:472) na Austrália, que relaciona como uso "o compartilhamento do entendimento entre decisores das políticas como base para o trabalho; o refinamento do conhecimento dos indivíduos sobre as questões do trabalho e a criação de um clima de expectativa para o desenvolvimento e implementação da política". Ou por Furubo (1994 *apud* WEISS, 1999:472), que demonstra que o governo sueco faz uso das avaliações, mas que eles se restringem à comunicação para a tomada de decisões, sem que os resultados venham a ser a base dessas decisões. No escritório de avaliação do Departamento de Educação dos Estados Unidos, a quantidade de citações de um dado estudo e as ações que são definidas em decorrência de determinados achados são os indicadores de uso (GINSBURG; RHETT, 2003). Não há, portanto, um único uso da avaliação e não se pode prever que uso será feito dela, independente dos usos originalmente propostos.

Seguindo o caminho percorrido pelos estudiosos de usos da avaliação, o presente estudo buscou inicialmente levantar a presença de elementos, no desenho avaliativo, que pudessem ser considerados uso-conducentes para, em seguida, buscar esses usos. A próxima subseção aproxima a categoria Utilidade do JCSEE, proposta para avaliação de programas, para a análise de políticas de avaliação.

2.4.2 A categoria Utilidade do JCSEE e sua adaptação para análise de políticas de avaliação.

Como mencionado na subseção dedicada aos padrões de qualidade da avaliação, a categoria Utilidade, para o JCSEE (texto na íntegra em http://www.wmich.edu/evalctr/jc/), é composta por sete indicadores, sintetizados na ilustração 04, que tratam de fatores que concedem à avaliação um padrão de qualidade em termos de sua utilização. Note-se que, ainda que se ocupe desses fatores, a categoria Utilidade do JCSEE não pressupõe uma consulta aos *stakeholders* sobre a sua percepção.

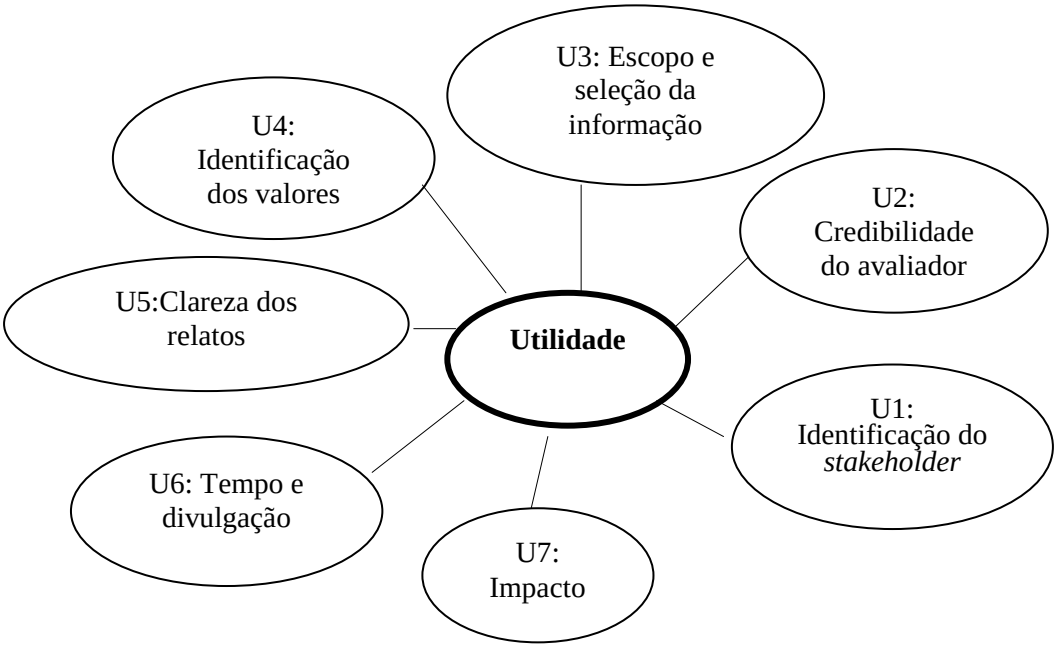


Ilustração 3:Representação dos sete indicadores do JCSEE da categoria Utilidade da Avaliação de Programa, utilizados pelo JCSEE.

É interessante perceber que essa categoria compreende elementos preditores de uso (discutido na próxima subseção), o que torna possível uma aplicação diagnóstica ou formativa e não apenas ao final de implementação da avaliação. De maneira geral, os sete indicadores que compõem a categoria Utilidade tratam de identificação dos *stakeholders*, seus valores e demandas avaliativas e, ao fazê-lo, preocupam-se com seu atendimento em tempo hábil e de forma clara para que se levante o impacto da avaliação. Aspectos de comunicabilidade e disseminação dos dados são abordados, bem como da credibilidade de quem avalia, como pode ser visto a seguir.

Categoria Utilidade da Avaliação de Programas

U1 – Identificação dos *stakeholders*. Pessoas envolvidas ou afetadas pela avaliação devem ser identificadas, de modo que suas necessidades possam ser atendidas.

U2 – Credibilidade do avaliador. Os indivíduos responsáveis pela avaliação devem ser confiáveis e competentes para a ação avaliativa, de forma que os achados da avaliação sejam críveis e aceitos ao máximo.

U3 - Escopo e seleção da informação. A informação coletada deve ser selecionada para atender questões pertinentes sobre o programa e ser efetiva quanto às necessidades e interesses dos clientes e dos *stakeholders* especificados.

> **U4 – Identificação de valores.** As perspectivas, procedimentos e *rationale* usados na interpretação dos resultados devem ser cuidadosamente descritos, de maneira a fundamentar os julgamentos de valor.
>
> **U5 – Clareza no relato da avaliação.** Os relatórios da avaliação devem descrever claramente o programa sob avaliação, incluindo seu contexto, seus objetivos, procedimentos e os achados da avaliação, de forma que os dados essenciais sejam fornecidos e que sejam de fácil compreensão.
>
> **U6 – Tempo e divulgação dos relatórios**. Relatórios parciais e finais devem ser apresentados aos usuários específicos em tempo hábil para sua utilização.
>
> **U7 – Impacto da avaliação.** As avaliações devem ser planejadas, implementadas e relatadas de jeito a incentivar continuidade pelos *stakeholders*, para que a probabilidade de sua utilização aumente. (JCSEE, 1999; tradução deste autor).

O JCSEE fala sobre as características de qualidade de um delineamento avaliativo para programas, mas não orienta sua aplicação em uma meta-avaliação. Baseado no JCSEE, Stufflebeam (1999) propôs um *checklist* para verificação da observância dos indicadores de qualidade. Nesse *checklist*, cada indicador recebe 10 itens de verificação e é pontuado em uma escala de adequação da avaliação aos padrões do JCSEE dividida em cinco níveis: 0-2 Fraco (*poor*), 3-4 Adequado (*fair*), 5-6 Bom (*good*), 7-8 Muito bom (*very good*) e 9-10 Excelente (*excellent*).

Stufflebeam (1999) recomenda que a experiência que obtiver um Fraco nos indicadores P1(Orientação para o Serviço), da categoria Propriedade, e A5 (Validade), A10 (Conclusões justificadas) ou A11 (Relato imparcial), na categoria Precisão, seja reprovada. Diferente da posição de Patton e seus seguidores sobre a obrigatoriedade do uso para a garantia da qualidade da avaliação, observa-se que nenhum dos indicadores da categoria Utilidade (marcados pela letra U) é determinante para a reprovação da experiência aos olhos do teórico que propõe o modelo mais influente de meta-avaliação de avaliação de programas. Uma hipótese para isso é que não há clareza, mesmo em 2009 (dez anos depois do *checklist*), do que realmente viabilize o uso de uma experiência avaliativa, especialmente quando consideradas as especificidades de cada contexto.

Nesta subseção, os sete indicadores do JCSEE para a categoria Utilidade são apresentados a partir dos itens de verificação propostos no *checklist* de Stufflebeam (1999, tradução deste autor). Em cada um deles, uma discussão é feita no sentido de aproximá-los às políticas de avaliação educacional, já que são, originalmente, voltados para a qualidade da avaliação de programas (não necessariamente políticas públicas), e do contexto brasileiro.

2.4.2.1 Itens de verificação para o indicador U1 – Identificação dos *stakeholders*.

- Identifica de maneira clara o cliente da avaliação.
- Envolve as lideranças na identificação de outros *stakeholders*.
- Levanta as necessidades de informação dos *stakeholders* potenciais.
- Utiliza *stakeholders* para identificação de outros *stakeholders*.
- Com o cliente, classifica os *stakeholders* pela sua importância relativa.
- Envolve os *stakeholders* durante o processo avaliativo.
- Mantém o delineamento aberto para servir novos *stakeholders*.
- Atende às necessidades avaliativas dos *stakeholders*.
- Atende uma gama adequada de *stakeholders* individuais.
- Atende uma gama adequada de organizações.

Os itens de verificação para o U1 podem ser divididos em dois grupos: a identificação e envolvimento dos *stakeholders* (sete primeiros itens de verificação) e o seu atendimento (três últimos itens de verificação).

Para aprofundamento dos sete primeiros itens, vale a pena fazer-se uma distinção entre o *stakeholder*, o cliente (*client*) e o usuário (*user, customer*) da avaliação. O *stakeholder* tem interesse ou responsabilidade direta sobre o objeto da avaliação, especialmente no nível decisório mais amplo. O cliente é responsável pela encomenda da avaliação e detém grande importância. Em alguns casos, como nas políticas de avaliação, a depender do contexto, do programa e das finalidades da avaliação, os demais *stakeholders* terão tanta importância quanto o cliente. Por fim, os usuários são aqueles a quem são dirigidos os elementos ou os resultados da avaliação; podem ser *stakeholders* principais, *stakeholders* com menor poder decisório, ou ainda indivíduos que fazem uso da avaliação de modo mais geral, sem relação específica com seu objeto (*users*), como no caso de uso informativo discutido na subseção anterior. O quadro a seguir apresenta os conceitos e um exemplo para esclarecer essa distinção, importante no momento de se fazer a hierarquia das demandas que resultarão nas perguntas avaliativas que, na seqüência, serão a base para a proposta de um determinado delineamento de avaliação.

Conceito	Definição	Exemplo
Stakeholder	Indivíduo, grupo de indivíduos ou organização com interesse direto sobre o objeto da avaliação. Suas demandas são a base para que se proponha o delineamento avaliativo.	Dirigentes escolares no caso de uma política de avaliação voltada para a escola pública.
Cliente	Indivíduo ou organização responsável pela encomenda (e compra) da avaliação. Normalmente, esse é o *stakeholder* principal.	Secretário de Estado que lança um edital para a "compra" de um serviço de avaliação.
Usuário	Indivíduo ou organização que utiliza, no todo ou em parte, os elementos da avaliação, do delineamento aos resultados; pode ou não ter um interesse direto no objeto sob avaliação e, nesse caso, pode ou não ser considerado um *stakeholder*.	*Stakeholder/usuário*: o dirigente escolar que, a partir dos dados da avaliação, vai utilizá-los em seu planejamento. Usuário-indireto: pesquisador interessado nas bases de dados construídas a partir de uma determinada política de avaliação.

Quadro 4: Definições de *stakeholder*, cliente e usuário da avaliação.

Para a meta-avaliação de políticas de avaliação, identificar o cliente é, sob um rápido olhar, aparentemente fácil: há um edital, uma chamada pública ou algum mecanismo de contratação da avaliação que, declaradamente, o define e a ele o avaliador se reporta em primeiro lugar. Isso não significa, entretanto, que não existam outros *stakeholders* principais. Por exemplo, em um ministério, o cliente pode ser uma dada diretoria, mas, no contexto, o *stakeholder* principal pode ser o ministro. Além disso, ministérios e secretarias não são blocos homogêneos. Em uma secretaria de educação, há outros *stakeholders* além do secretário, como as várias diretorias ou superintendências, as coordenações de programas específicos, as representações regionais e as comunidades escolares, que podem ter agendas políticas e demandas avaliativas muito diferentes, por vezes conflitantes. Acresce-se a isso o fato de que, no Brasil, as esferas administrativas (federal, estadual e municipal) trabalham de maneira autônoma, ainda que em colaboração, definida em Lei (Constituição de 1988; Lei 9394/96). É muito freqüente haver conflito entre as esferas quanto às políticas de avaliação. Um exemplo disso é a submissão (ou não) das universidades estaduais ao SINAES. Nem todas aderiram ao Sistema vez que a proposta é federal e elas estão na esfera estadual.

Os *stakeholders* e usuários, nas políticas de avaliação educacional no Brasil, podem estar nos seguintes níveis: político, técnico centralizado, técnico descentralizado e não técnico, além de haver um grupo de usuários não diretos, como pode ser visto no quadro a seguir. A centralização, nesse caso, refere-se à alocação do indivíduo no órgão central (como o ministério ou as secretarias de educação).

Tipo	Nível	Representantes	Exemplos de demandas avaliativas
Stakeholders	Político	Ministros, secretários, superintendentes.	Macro; voltadas para decisões que afetam todo o sistema.
	Técnico central	Burocracia média nos órgãos centrais de governo, como secretarias, diretorias, superintendências.	Macro, voltadas para o sistema, e micro, voltadas para *accountability* e regulação desse sistema.
	Técnico não central	Burocracia média nas representações regionais e nas escolas (diretores regionais, dirigentes escolares, professores e coordenadores).	Micro, voltadas para a gestão local e para o desenvolvimento do trabalho em sala de aula.
	Não técnico	Pais e alunos	Macro, em termos gerais, e micro, em termos do local onde a política se concretiza para cada aluno (e, em decorrência, para seus pais).
Usuários	Não diretos	Sociedade em geral; academia em geral; pesquisadores no campo da avaliação; instâncias políticas outras que não as diretamente envolvidas com o objeto, dentre uma infinidade de possibilidades.	Geral, em termos globais, ou específicas, como interesse na base de dados, na metodologia, no relato de estratégias de comunicação, etc.

Quadro 5: Síntese dos níveis de *stakeholders* e usuários para as políticas de avaliação educacional

Em relação à análise do atendimento da demanda, quando o objeto da avaliação é uma política de avaliação educacional, é preciso refletir sobre o que seja esse atendimento. É Stufflebeam (1974) quem, no início da década de 70, chama atenção que os interesses desses *stakeholders* são muito diferentes e que, provavelmente, um único delineamento avaliativo não dará conta de atendê-los. Isso é especialmente importante porque, de maneira geral, os editais para seleção de propostas de avaliação – meio pelo qual a política é concretizada - são vagos, por vezes decorrentes de alguma exigência de órgãos financiadores sem uma demanda real do ministério, da secretaria de Estado ou de município, o que implica um trabalho preliminar do avaliador em definir as reais necessidades avaliativas para, em seguida, propor como atendê-las. Rebolloso *et alii* (2002) lembram que a dimensão política e a necessidade de acomodação dos valores postos têm despertado a consciência, no campo da avaliação, sobre a importância do processo de negociação e da busca de consenso entre os atores, no que são reforçados pelo posicionamento de Patton quanto ao envolvimento dos *stakeholders*. Shula e Cousins (1997) ressaltam que é central saber se os *stakeholders* concordam sobre os objetivos do programa e sobre suas finalidades. Entretanto, diante do panorama de diversidade de *stakeholders* apresentado nos quadros acima, não se deve esperar consenso.

Por outro lado, quando o foco da meta-avaliação volta-se para o atendimento de demandas sociais, concorda-se com Zákia Souza (em entrevista a YAZBECK, 2007:17) de que não tem havido procedimentos voltados para seu levantamento como etapa das sistemáticas de avaliação. Além disso, há que considerar que, no Brasil, pais e alunos, *stakeholders* centrais vez que é para eles

que as políticas educacionais são implementadas, não se incluem, de maneira geral, como demandantes de informação da avaliação, ainda que, após o Provão, essa característica tenha começado a mudar pelo menos no Ensino Superior.

Em relação ao atendimento das demandas dos usuários, Luis Carlos Freitas, ao criticar o Estado Avaliador brasileiro, lembra que é gerada uma quantidade enorme de dados que o usuário da avaliação (mesmo aquele ligado ao sistema educacional) não consegue consumir (em entrevista a YAZBECK, 2007:18). Encharcar o usuário com dados não significa atendê-lo. Já Abramowicz atribui ao fato de normalmente os usuários não serem ouvidos em relação a suas demandas avaliativas um fator de impedimento para a utilização da avaliação (ABRAMOWICZ, 1994).

Por fim, para análise do atendimento da demanda, um desafio se coloca: estabelecer a linha que separa os objetivos da política de avaliação educacional daqueles do programa educacional (ou do sistema educacional) sob avaliação. Há uma tendência para objetivos gerais muito amplos, comuns a várias políticas, dentre as quais a de avaliação, e de, em conseqüência, esperar-se que o atendimento da demanda da avaliação seja, em verdade, o atendimento da demanda para o programa educacional sob avaliação (ver discussão sobre finalidades na subseção anterior). Nesse panorama, a avaliação jamais será capaz de atingir, isoladamente, a finalidade determinada para a política maior.

Se, como posto por Locatelli (2001), é na escola que a mudança ocorre em termos de qualidade educacional, não necessariamente a escola é *stakeholder/usuário* para a avaliação: o objeto pode ser o sistema educacional e a demanda avaliativa relacionada ao levantamento de informações sobre como e quanto esse sistema favorece a ocorrência da mudança. A escola, nesse delineamento, é fonte de dados. O sistema educacional é o objeto sob a avaliação. Devolver os dados dessa avaliação para a escola é inócuo, vez que ela não está sendo avaliada. Esse foi o principal problema com o SAEB: tendo sido desenhado para atender aos níveis políticos federal e estaduais, foi duramente criticado por não oferecer dados desagregados por (e para) a escola e, pior ainda, por a escola não conseguir utilizar os seus resultados.

Identificar os *stakeholders* e suas demandas para avaliação e, em um segundo momento, analisar o atendimento dessas demandas constituem-se no primeiro indicador da categoria Utilidade. Como esses *stakeholders* percebem (e recebem) a avaliação está diretamente relacionado ao segundo indicador, Credibilidade, tratado na subseção a seguir.

2.4.2.2 Itens de verificação para o indicador U2 – Credibilidade do avaliador

- Emprega[44] avaliadores competentes.
- Emprega avaliadores nos quais os *stakeholders* confiam.
- Emprega avaliadores que podem responder a preocupações dos *stakeholders*.
- Emprega avaliadores que adequadamente respondem a questões de gênero, *status* socioeconômico, raça, e diferenças culturais e de linguagem.
- Assegura que o plano de avaliação atenda às principais preocupações dos *stakeholders*.
- Ajuda os *stakeholders* a entenderem o plano de avaliação.
- Fornece aos *stakeholders* informações sobre aspectos de qualidade técnica e operacional do plano de avaliação.
- Responde adequadamente às críticas e sugestões dos *stakeholders*.
- Mantém-se a par das forças políticas e sociais.
- Mantém as partes interessadas informadas sobre o progresso da avaliação.

O indicador Credibilidade é de suma importância quando da análise da utilidade de um delineamento de avaliação porque é sabido que aqueles que não crêem na avaliação e/ou no avaliador/equipe avaliadora não utilizarão seus resultados. Sob pressão, poderão até fazê-lo, mas de modo ritualizado, apenas para cumprir com a obrigação. Os dez itens de verificação do indicador Credibilidade, propostos por Stufflebeam (1999), podem ser distribuídos em relação à competência do avaliador ou da equipe avaliadora (quatro) e em relação ao comportamento desses indivíduos (seis).

Não há disputa sobre a importância do emprego de uma equipe de avaliação competente, crível e sensível às questões de raça, gênero, condição sócio-econômica, seja na avaliação primária, seja na meta-avaliação, como discutido na subseção que tratou da formação da equipe de meta-avaliadores. Entretanto, em grande maioria, os *stakeholders* não têm condições de analisar tecnicamente essa competência. Uma estratégia adotada pelos *stakeholders* é confiar na percepção geral sobre a integridade profissional da equipe de avaliadores, em lugar da competência técnica. Outra estratégia utilizada pelos *stakeholders* para analisar a credibilidade de uma dada equipe de avaliadores é transferir sua percepção para a instituição de vínculo da equipe. Por exemplo, se uma equipe de avaliadores é ligada a uma determinada universidade de prestígio, é muito provável que os avaliadores vinculados a ela sejam competentes. Além disso, como discutido por Davok (2006:89), "de maneira geral, os *stakeholders* depositam maior confiança em equipes externas de avaliação, por sua independência com o objeto avaliado e pela objetividade que essa

44 O termo em inglês é *engaje*, no sentido de comprometer, envolver, empregar. Optou-se por "empregar" por se estar lidando com situações de trabalho.

condição pode dar ao relatório final da avaliação". Em relação às políticas de avaliação educacionais no Brasil, um efeito colateral da escolha de avaliadores externos ligados às universidades é observado nas escolas públicas sujeitas à avaliação: para elas, os avaliadores, apesar de competentes, não conhecem sua realidade. Nesse sentido, não há descrença quanto à capacidade técnica, mas há descrédito na capacidade de entendimento do objeto avaliado, no caso a educação básica pública, às vezes confirmada por relatórios técnicos escritos em jargão acadêmico. Diante dessa percepção, o item verificador "Mantém-se a par das forças políticas e sociais" poderia ser enriquecido por "Mantém-se a par das forças políticas e sociais e da realidade do objeto avaliado".

De acordo com Stufflebeam (1974), o indicador Credibilidade pergunta se o(s) público(s) da avaliação acredita(m) no avaliador e se pensa(m) que o processo avaliativo está livre de vieses. Falando em política de avaliação em larga escala, essa categoria está voltada para a isenção/independência da avaliação ou, dito de outra maneira, de seu distanciamento das questões político-partidárias, de maneira que os resultados obtidos sejam imparciais.

Em relação aos itens verificadores de comportamento do avaliador /equipe de avaliação, o primeiro deles (Assegura que o plano de avaliação atende às principais preocupações dos *stakeholders*) em parte repete os itens verificadores de levantamento da demanda dos *stakeholders* relativos ao indicador U1. Essa questão remete o leitor à crítica apresentada por Widmer (2007), relatada na seção que trata de padrões de qualidade da avaliação, ao buscar aplicar os padrões do JCSEE na Suíça.

Os três itens seguintes falam da relação avaliador x *stakeholders*. No caso de políticas de avaliação em larga escala, esses itens levantam o desafio da comunicabilidade. O meta-avaliador precisa buscar as estratégias e os canais estabelecidos entre a equipe de avaliação e os *stakeholders* tanto para o esclarecimento dos planos, dos processos e dos relatórios, como também no recebimento e análise das críticas e sugestões apresentadas. Já o item "Mantém as partes interessadas sobre os progressos da avaliação" remete às posições de Davok (2006) e Penna Firme e Letichevsky (2002). Segundo essas autoras, a credibilidade é conferida pela transparência nos processos e nos resultados (inclusive com a discussão de seus limites).

Em outra corrente, Patton define credibilidade como "um conceito complexo que inclui a percepção de precisão, justiça, e confiança da avaliação" (1997:250). Cooksy e Caracelli (2005:35-36) operacionalizaram esse conceito a partir do grau de inferência a ser feito nas

relações de causa x efeito estabelecidas pelas avaliações, observadas a partir de um modelo lógico no qual foram buscadas as evidências para tais relações. O exemplo que oferecem é ilustrativo: uma relação de baixa inferência é vista, por exemplo, entre “novas tecnologias disponibilizadas” e “24 cultivares de arroz” oriundas de pesquisa em agricultura. Já o impacto dessas cultivares no aumento da renda do trabalhador rural não é tão claro, o que requer um alto nível de inferência. Para esses autores, o nível de inferência só faz sentido se apoiado por evidências. A credibilidade é baixa se, qualquer que seja o nível de inferência, não há evidências que suportem a relação causa x efeito. A credibilidade cresce à medida que essas evidências são apresentadas. Esse modelo lógico permite identificar a credibilidade qualquer que seja o delineamento metodológico adotado (COOKSY; CARACELLI, 2005). Tal abordagem pode ser usada para analisar a credibilidade da avaliação, mas deixa de fora a credibilidade da equipe avaliadora. Talvez seja interessante, para um próximo modelo, considerar as duas abordagens.

Uma vez identificados os *stakeholders* e suas demandas e observada a credibilidade dos avaliadores, o próximo indicador da qualidade da avaliação na categoria Utilidade é o Escopo Avaliativo. É sobre isso que trata a subseção a seguir.

2.4.2.3 Itens de verificação para o indicador U3 - Escopo e seleção da informação.

- Entende os requisitos da avaliação mais importantes para o cliente.
- Entrevista *stakeholders* para determinar suas perspectivas.
- Assegura negociação entre avaliador e cliente sobre públicos pertinentes, questões avaliativas e demanda de informação.
- Atribui prioridade para os *stakeholders* mais importantes.
- Atribui prioridade para as questões mais importantes.
- Flexibiliza a adoção de novas questões durante a avaliação.
- Obtém informação suficiente para atender às questões mais importantes dos *stakeholders*.
- Obtém informação suficiente para avaliar o mérito do programa.
- Obtém informação suficiente para avaliar o valor do programa.
- Aloca esforços avaliativos de acordo com as prioridades determinadas às informações necessárias.

Os itens verificadores do indicador U3, como no caso dos anteriores, podem ser percebidos em dois grupos: os seis primeiros falam da priorização, pelo avaliador, da demanda levantada junto aos *stakeholders* para a definição do escopo sob avaliação e, no segundo conjunto, da obtenção da informação em quantidade suficiente em respeito a esse escopo. Mais uma vez, percebe-se uma sobreposição dos itens verificadores do U3 e do U1, especialmente quanto às demandas

avaliativas, embora o U3 se volte para aquilo que realmente é concretizado, enquanto que o U1 trata mais notadamente da demanda.

Estabelecer o recorte da avaliação, que em políticas de avaliação é comumente referido como definição de variáveis ou desenho da matriz de referência, é, talvez, isoladamente, a etapa de maior impacto político nessa atividade. O simples fato de incluir um ou outro aspecto na matriz dá ao aspecto escolhido uma condição de importância sobre aquilo que não foi considerado, independente do fato de essa escolha estar também condicionada a questões de técnica, de custo e de tempo e não necessariamente a uma hierarquia de importâncias. Além disso, há uma contradição posta na origem: o aumento da amplitude do escopo avaliativo implica a perda de profundidade, especialmente para avaliações em larga escala. Por exemplo, em termos de testes, quanto maior o número de variáveis, tanto menor o número de questões de um teste para avaliá-las individualmente, já que a capacidade do respondente não é elástica.

Os itens verificadores do indicador U3 (para avaliação de programas) são todos utilizáveis para políticas de avaliação. Entretanto, há alguns cuidados a tomar. O primeiro deles é o cruzamento dos dados de prioridade do cliente e dos *stakeholders* nos momentos da encomenda avaliativa (e de sua negociação) e da entrega final. As pressões políticas e operacionais são muito fortes à medida que se dá a implementação de uma política de avaliação e, por vezes, chegam a desfigurá-la. Durante o processo, é possível inclusive que o cliente – uma determinada gestão governamental – seja mudado e a nova gestão, ainda que presa aos termos do edital de seleção, tenha outras prioridades ou mesmo posicionamentos ideológicos contrários à avaliação contratada (ABRAMOWICZ, 1994, WEISS, 1998; 1999). Por outro lado, é necessário considerar uma defesa teórica para o recorte das demandas avaliativas que resultarão no escopo da avaliação, sob pena de se obter um delineamento avaliativo frágil ou não defensável.

O segundo cuidado para os meta-avaliadores é observar a suficiência da informação. Em alguns casos, vê-se uma coleta de dados superior à necessidade de resposta às questões prioritárias, até pela possibilidade da mudança das mesmas no processo. Dados demais, no entanto, não contribuem para a qualidade da avaliação, chegando, em alguns casos, a atrapalhar. Esse é o caso, por exemplo, de longos questionários aplicados em uma determinada fonte (como os dirigentes escolares) que, por interferir com o seu trabalho rotineiro, acabam por tornar o “custo de contribuir com informações” superior ao “benefício da entrega do relatório final”. Dados demais coletados afetam o último item verificador desse indicador: os esforços avaliativos, longe de

serem concentrados nas prioridades, dissipam-se na coleta e tratamento de dados não diretamente interessantes para os *stakeholders* principais.

Um terceiro cuidado diz respeito ao delineamento da coleta de dados, especialmente no caso das avaliações em larga escala, em termos das amostragens. A informação deve ser suficiente, representativa e também permitir que suas análises contribuam para o melhoramento do objeto da avaliação, quanto ao mérito e/ou ao valor. Ginsburg e Rhett (2003), por exemplo, ao falar sobre delineamentos experimentais, lembram que habitualmente os desenhos amostrais randômicos consideram o grupo controle e o grupo impactado, mas raramente observam características particulares de um dado programa, o que inviabiliza o provimento de orientações sobre como melhorá-lo.

O conceito de avaliação que é adotado no presente trabalho fala sobre o estabelecimento de um juízo de qualidade sobre dados relevantes, tendo em vista uma tomada de decisão, conforme Luckesi (2000). O indicador U3 trata do recorte dos dados relevantes, considerados os *stakeholders* da avaliação. O U4 discute os elementos que definirão os critérios que permitirão o juízo de qualidade ou, como o JCSEE prefere, o julgamento de valor.

2.4.2.4 Itens de verificação para o indicador U4 – Identificação de valores.

- Considera fontes alternativas de valores para interpretação dos achados da avaliação.
- Fornece uma base clara e defensável para os julgamentos de valor.
- Determina a parte apropriada para fazer as interpretações valorativas.
- Identifica necessidades sociais pertinentes.
- Identifica necessidades pertinentes dos usuários.
- Leva em consideração as leis pertinentes.
- Leva em consideração a missão institucional.
- Leva em consideração os objetivos do programa.
- Leva em consideração os valores dos *stakeholders*.
- Apresenta interpretações alternativas fundamentadas em base valorativa crível, ainda que conflitante.

Como no caso dos indicadores anteriores, alguns itens de verificação estão sobrepostos a outros já discutidos (Identifica necessidades sociais e Identifica necessidades dos usuários pertinentes) e também este indicador pode ser analisado a partir de dois grupos: a identificação das bases para a definição dos valores e a determinação dos valores em si. Vale destacar que, nesse indicador, o JCSEE não está lidando com os conceitos de valor e de mérito discutidos por Stufflebeam (1997),

mas, mais restritamente, refere-se às bases para que o julgamento do objeto seja informado no processo da avaliação.

Para definição dos valores, os itens consideram as leis, a missão institucional, os objetivos do programa e os valores dos *stakeholders*. No caso de avaliação de políticas de avaliação, é essencial o respeito às leis (já discutido na subseção que tratou do delineamento da avaliação) e às linhas políticas de Estado (com recomendável afastamento das questões político-partidárias). Não é tão simples considerar os objetivos do programa, por todas as questões já postas quanto à amplitude dos mesmos e pelas posições conflitantes dos *stakeholders*.

A valoração é feita a partir da identificação de patamar aceitável e da distinção entre ele e o que não é aceitável em termos do objeto da avaliação. Ou, no caso da avaliação educacional, do que separa o nível de qualidade mínimo aceitável daquilo que não é possível aceitar. Esse é o caso da definição da linha de corte, por exemplo, em avaliações quantitativas, e da identificação dos elementos-chave para que a experiência seja considerada bem sucedida, em delineamentos qualitativos. No Brasil, a análise das políticas de avaliação mostra que a definição do escopo não tem sido um problema, mas que a determinação do valor não é freqüente. Esse fato contribui para a dificuldade de interpretação dos resultados. O SAEB, por exemplo, apresenta seus resultados em uma escala contínua de desempenho da primeira até a última série. Inicialmente, procurava informar onde na escala deveria ser situado o desempenho ideal de cada uma das séries abordadas pela política. Por problemas técnicos (defender a escala contínua sem uma aplicação de provas em todas as séries é muito difícil), o SAEB optou apenas por divulgar a escala e por informar o percentual de alunos em cada nível, como anteriormente tinha feito a Bahia, mas nas escalas por série. Essa opção faz com que os *stakeholders* saibam o desempenho dos Estados nas séries avaliadas, mas não saibam se esse desempenho é o esperado (e aceitável) ou não. O Provão (ENC) utilizava uma abordagem referenciada à norma e permitia, dentre aqueles cursos avaliados, um posicionamento relativo sobre a qualidade encontrada, sem jamais ter definido o patamar mínimo de qualidade aceito. O mesmo ocorre hoje com o ENADE (VERHINE, DANTAS, 2006). Das políticas de avaliação federais no Brasil, a única em larga escala que não deixa dúvidas quanto à valoração é a adotada pela CAPES para a avaliação da pós-graduação. Qualquer estudante de pós *stricto sensu* sabe (o Sistema CAPES, originado na década de 70, é consolidado) que se seu curso não obtiver Conceito 3 na avaliação trienal, deixará de ser recomendado, o que significa que demonstrou um nível não aceitável de qualidade.

O mais interessante dos itens de verificação do U4 é a determinação da parte apropriada para fazer as interpretações valorativas. Em avaliações externas, é comum ao avaliador propor e conduzir os processos que resultem na base para o julgamento de valor, a contribuir para a tomada de decisão feita pelo gestor. Por outro lado, uma vez disseminadas as informações oriundas da avaliação, não há como impedir que os *stakeholders* e usuários, independente de quaisquer bases para o julgamento, façam suas próprias interpretações e adotem posicionamentos. Nesse sentido, cabe ao meta-avaliador analisar se a experiência avaliativa adotou medidas para preservar as fontes, esclarecer as bases valorativas, informar as limitações da avaliação e ressaltar que essas bases valorativas são adequadas a um determinado escopo de avaliação, mas não são generalizáveis.

Na lógica dos indicadores da Categoria Utilidade, uma vez garantidos os elementos para determinação do julgamento de valor, os dois próximos itens abordam a questão da comunicação entre avaliadores e avaliados. Grande número de estudiosos sobre utilização da avaliação reputa à comunicação o papel mais importante para assegurar que os usos sejam feitos, inclusive extrapolando os usos instrumentais para a questão da influência (LAWRENZ; GULLICKSON; TOAL, 2007).

2.4.2.5 Itens de verificação para o indicador U5 – Clareza no relato da avaliação

- Relata de maneira clara as informações essenciais.
- Divulga relatórios breves, simples e diretos.
- Focaliza relato das questões contratuais.
- Descreve o programa e seu contexto.
- Descreve os propósitos da avaliação, seus procedimentos e achados.
- Fundamenta conclusões e recomendações.
- Evita utilização de jargão técnico.
- Utiliza a linguagem dos *stakeholders* nos relatos.
- Fornece sumário executivo.
- Fornece relatório técnico.

Os itens verificadores do indicador U5 – Clareza no relato da avaliação – abordam a forma e a escolha do conteúdo dos relatórios (ou dos relatos) da avaliação de programa. Quanto à forma, a meta-avaliação deve observar a adequação ao(s) público(s)-alvo da linguagem utilizada pelos avaliadores (sem jargão) e a composição de relatórios claros, breves e diretos (parágrafos curtos são sempre bem vindos). Quanto ao conteúdo, o último item aponta para a necessidade de uma avaliação de qualidade ter um relatório técnico diverso do relatório de resultados, mas, em ambos os casos, o meta-avaliador deve observar se há um sumário executivo, se as questões contratuais estão endereçadas, se houve um relato de “propósitos, procedimentos e achados” e se as

conclusões e recomendações estão fundamentadas (o que contribui enormemente para a credibilidade, como visto no U2).

Esses itens podem facilmente ser aplicados na meta-avaliação de políticas de avaliação, mas alguma adequação precisa ser feita. Em termos de forma, é preciso considerar a diversidade dos *stakeholders*, não só entre os grupos, mas dentro de cada grupo (MAY, 2004). Tome-se o exemplo do SAEB: um dos *stakeholders* principais é o governo de cada estado. Há estados que dominam a linguagem da avaliação porque, de algum modo, antecedem o próprio SAEB, como é o caso de São Paulo; há estados, como o Amapá, que entraram em contato com as políticas de avaliação pelas mãos do SAEB. Os cuidados com a forma de apresentação dos dados nesse segundo caso são muito maiores que no primeiro caso, sendo justificada uma abordagem de alguma maneira didática. Infelizmente, os relatórios são únicos (mesmo que, desde 2005, customizados por Estado para dificultar *ranking*s) e há ainda muita queixa de que a linguagem é bastante distante daquela para "leigos". A Prova Brasil, com resultados individuais por escola, teria que ter relatórios personalizados, mas extremamente breves, com uma capacidade de comunicação de massa, já que as escolas precisariam discutir seus resultados com pais, alunos e professores. As duas ilustrações a seguir demonstram esse esforço de comunicação: o relatório da Prova Brasil é colorido, facilmente transformado em cartaz, e possibilita comunicação imediata com os diversos públicos da escola avaliada. O problema nesse caso é o oposto do SAEB: o relatório não dá nenhuma pista sobre a qualidade do desempenho; apenas apresenta seus dados.

Ilustração 4: Frente do Relatório da Prova Brasil. http://sistemasprovabrasil2.inep.gov.br/ProvaBrasil/2005/BA/29191327.pdf. Nome da escola retirado.

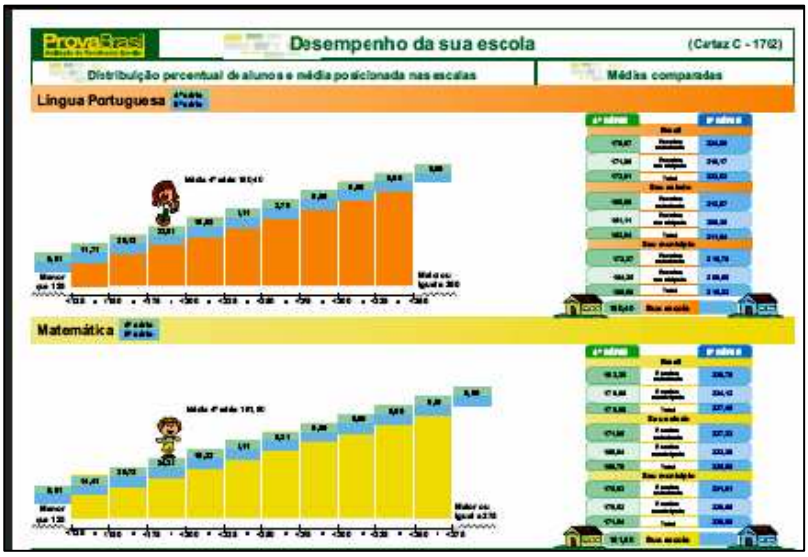


Ilustração 5: Verso do relatório da Prova Brasil. Disponível em http://sistemasprovabrasil2.inep.gov.br/ProvaBrasil/2005/BA/29191327.pdf. Nome da escola retirado.

O esforço de comunicação também é observado no ENADE: o resultado da prova do estudante é acessado por meio de senha individual na *internet* e restringe-se a uma página-síntese. Já o Relatório do ENADE (mais técnico e geral), para os públicos dos cursos de graduação e instituições de nível superior, traz detalhadamente as finalidades do exame no contexto SINAES, os resultados gerais de desempenho, os resultados do questionário sócio-econômico, além de uma breve descrição da metodologia (ainda que esse capítulo seja de difícil compreensão para leigos).

O problema do capítulo da metodologia no relatório do ENADE é bastante comum aos relatos de avaliação: é muito difícil o equilíbrio entre a linguagem técnica e a linguagem leiga, de modo a, por um lado, ter-se um relatório de fácil leitura, mas, por outro lado, o relatório não ser "infantilizado", de certo modo demonstrando desrespeito aos públicos mais leigos. Em especial, é longo o caminho entre a linguagem da estatística e a linguagem utilizada pelos educadores. Nesse sentido, May (2004) argumenta que para se entender um dado não é necessário o conhecimento sobre o método estatístico utilizado para obtê-lo e que, para favorecer a interpretação do dado, a informação estatística pode ser disponibilizada em linguagem familiar. Em alguns relatórios de avaliação, o meta-avaliador poderá encontrar pequenos "glossários", por vezes exemplificados.

Ainda em relação à forma dos relatos, Alkin (*apud* ABRAMOWICZ, 1994:91) lembra que o tipo de informação pode ser quantitativo ou qualitativo e que há uma multiplicidade de mídias para veiculá-las, tais como o relatório impresso, a apresentação oral, em fotografia, filme, vídeos, etc. Essa consideração é especialmente importante no caso de o avaliador precisar esclarecer ao

stakeholder sobre os dados e achados quando, em larga escala, o contato pessoal com tantos representantes é difícil ou mesmo impossível. Mais recentemente, no Brasil, a vídeo conferência tem sido utilizada tanto nas discussões para o delineamento da avaliação como no debate sobre os resultados. Já Lawrenz *et alii* (2007) advogam em favor de brochuras, boletins, sumários curtos com gráficos e tabelas e memos com síntese de resultados como formato de relato avaliativo, especialmente considerado o pouco tempo de divulgação (a ser visto no indicador U6).

Quanto ao conteúdo dos relatos de avaliação, recorre-se mais uma vez ao artigo de May (2004) na observação de um texto ideal: deve-se considerar cada experiência de pesquisa como um processo contínuo que tem uma pergunta e uma conclusão, um começo e um final; e que o produto de tal processo é mais útil quando o círculo completo é fechado e o começo e o final são conectados. O mesmo se aplica à avaliação. O meta-avaliador, em seu trabalho, precisa identificar nos relatos avaliativos "o começo e o final", o propósito, as finalidades, os processos e, finalmente, os produtos da avaliação. Dada à diversidade de *stakeholders* e usuários nas políticas de avaliação educacional, não há um único formato de relatório (ou de relato), mas formatos diversos com níveis diferentes de profundidade e detalhamento de processo e produtos dirigidos aos vários públicos. A definição do grau de detalhamento da informação a ser encaminhada a cada público tem clara natureza política. É também comum que críticas sobre a decisão da informação demonstrem receio quanto à "manipulação de dados".

É importante, na aplicação do indicador U5 para a análise de políticas de avaliação, que seja acrescido o item *Relata os limites da avaliação.* Este item refere-se, por exemplo, à divulgação das margens de erro (que a população se acostumou a lidar a partir das pesquisas de opinião em tempos de eleição), ou a esclarecimentos sobre o escopo avaliativo (U3), entendido como um recorte do objeto avaliado. Isso é especialmente importante no caso de avaliações que resultam em classificações e *ranking*s: é preciso que o meta-avaliador observe como os *rankings* foram apresentados, para evitar injustiças para os avaliados.

Discutidos os itens de verificação quanto à clareza do relato, há ainda outros aspectos da comunicação de grande impacto para a utilização (ou não) da avaliação. O indicador U6 trata desse tema e é detalhado na subseção a seguir.

2.4.2.6 Itens de verificação para o indicador U6 – Tempo e divulgação dos relatórios

- Apresenta relatórios parciais aos usuários-alvo.
- Entrega o relatório final quando é necessário.

- Interage em tempo conveniente com os formuladores do programa.
- Interage em tempo conveniente com o *staff* do programa.
- Interage em tempo conveniente com os usuários do programa.
- Interage em tempo conveniente com os públicos interessados.
- Interage em tempo conveniente com os meios de comunicação.
- Emprega mídia adequada para alcançar e informar os diferentes públicos.
- Mantém breves as apresentações.
- Utiliza exemplos para ajudar os públicos a relacionar os achados com situações práticas.

Também nesse indicador há sobreposição de alguns itens em relação ao indicador anterior, embora o U5 trate da forma e do conteúdo do relato e o U6 se volte para quando e como esse relato atinge os *stakeholders* e usuários. Os itens de verificação do U6 podem ser observados exatamente nessa divisão: cinco deles tratam da conveniência de tempo da comunicação e os cinco restantes falam das estratégias utilizadas para que a comunicação alcance seu público.

Como discutido na seção sobre avaliação, os resultados de processos avaliativos são perecíveis. Por exemplo, em casos de avaliações conduzidas para que o decisor escolha se continua ou não com as ações em um determinado programa, os resultados perdem completamente sua finalidade quando são devolvidos ao público em um momento posterior àquele da tomada de decisão. Da mesma maneira, se a forma de divulgação não privilegia informações que levem à formação de um juízo de qualidade sobre o objeto avaliado, pode resultar em um resultado “oco” de significado. Essa categoria, portanto, trata de uma análise das condições básicas para a utilização dos resultados: entrega dos resultados no tempo certo e a maneira utilizada para que o(s) público(s) consiga(m) se relacionar com eles. Concorda-se com Weiss (1999) quando a autora argumenta que a disseminação das informações não é passatempo, mas uma atividade que requer tempo, raciocínio e energia para que as pessoas certas sejam atingidas.

Como ocorre nos indicadores anteriores, os itens de verificação são facilmente transpostos para uma meta-avaliação de políticas de avaliação, com uma ressalva: a definição do que sejam os formuladores, o *staff* e os usuários do programa. Na transposição para a avaliação de políticas de avaliação, esse item refere-se à própria avaliação. Tem-se, portanto, os clientes, os *stakeholders* e os usuários da avaliação. O *staff* é a própria equipe formuladora/implementadora da política de avaliação.

Um dos aspectos a ressaltar diz respeito à conveniência dos tempos em relação à hierarquia e à priorização das demandas dos *stakeholders*, cruzando, portanto o U1 com o U6. Dito de outra forma, o cliente e os *stakeholders* principais são os primeiros a serem informados e só então os

outros níveis hierárquicos têm acesso às informações. Esse cuidado deve ser observado tanto pelas políticas federais, quanto pelas estaduais e municipais, especialmente porque suas representações maiores, tão logo a informação esteja disseminada, serão chamadas para explicá-la e discuti-la. "Emboscar" os *stakeholders* principais com o lançamento de informações sem que os mesmos tenham conhecimento prévio pode criar uma resistência à política de avaliação que tornará seu uso mais difícil.

Um outro aspecto a levar em consideração diz respeito à diversidade da conveniência dos tempos quando são analisados todos os *stakeholders* envolvidos. Por exemplo, para uma política de avaliação que alimente a tomada de decisão dos órgãos centrais de educação, como ministérios e secretarias, as informações precisam estar disponíveis antes do momento de planejamento orçamentário do ano subseqüente. Caso contrário, quaisquer decisões para o melhoramento do objeto sob avaliação precisarão esperar até o ano seguinte, já que há sérias amarras das ações ao orçamento. Já no caso de escolas públicas como *stakeholders*, o tempo de informação é aquele anterior à jornada pedagógica, na qual são planejados os cursos para o ano subseqüente. Nem sempre o planejamento orçamentário dos órgãos centrais coincide com as jornadas pedagógicas. Políticas de avaliação estaduais implementadas na Bahia e no Paraná, por exemplo, resolveram essa questão ao entregar à escola o tratamento e a análise de dados (ver BONAMINO; BESSA; FRANCO, 2004). Para análises simples de dados, esse procedimento é aceitável. Para análises que impliquem utilização de abordagens de pesquisa mais complexas, como as políticas que buscam comparabilidade ao longo do tempo, essa estratégia é danosa por comprometer a validade dos resultados. Quando considerado o *staff* da política de avaliação, a informação deve chegar a tempo de permitir que essa equipe corrija os rumos de implementação para o atingimento dos objetivos da política, em um *feedback* constante.

Políticas de avaliação educacional em larga escala, como o SAEB, dada sua complexidade, normalmente divulgam resultados pelo menos um ano após a coleta de dados. Esse tempo, ainda que curto para os técnicos que processam os dados, é longo demais para os *stakeholders* que demandam tais informações. Esse paradoxo é discutido, dentre outros, por Lawrenz, Gullickson e Toal (2007), especialmente porque é árdua a busca do equilíbrio entre a necessidade de entrega da informação devidamente tratada e analisada, por um lado, e de entrega de informação em tempo que favoreça a utilização. Nessa busca, frequentemente elementos de precisão são perdidos. A equipe de avaliação precisa determinar qual o nível de precisão minimamente aceitável. O tratamento e a análise dos dados coletados e a transposição para relatórios demandam tempo, mas são cruciais para a qualidade da comunicação.

Ainda sobre tempo e perecibilidade da avaliação, é importante registrar que, como posto por Ferrer (1997), o fluxo de informações na experiência avaliativa é contínuo e não deve ser considerado apenas em relação aos relatórios finais, mas também aos parciais e ao projeto avaliativo. Quanto mais os *stakeholders* ficam informados sobre a avaliação, da concepção ao fechamento, tanto mais provável a incorporação dos conceitos e posterior uso dos resultados.

Um tópico a refletir é a interação política de avaliação - meios de comunicação, em especial a imprensa. É possível hoje acompanhar ações avaliativas pela *internet*, através de *sites* especializados, como o faz o INEP (ainda que, a partir de 2004, tenha havido pouquíssima atualização em relação a aspectos técnicos da grande parte das avaliações implementadas – ver www.inep.gov.br – posição em fevereiro de 2009). Entretanto, o hábito de consulta a boletins *on-line* ainda não está estabelecido e muitas unidades escolares, ainda que tenham computadores, não contam com acesso à rede. No Estado da Bahia, um meio muito eficaz de atingir as escolas estaduais é o Diário Oficial do Estado, recebido diretamente em muitas delas e, em outros casos, consultados nas DIREC. As redes municipais, especialmente pela quantidade de escolas na zona rural, têm o acesso mais difícil – até 2004 ainda havia escolas na rede sem eletricidade e não havia entrega regular de jornais em alguns municípios menores. Em termos da TV, pode-se afirmar que seria bem mais eficaz no atingimento do conjunto de *stakeholders*. A ponderação que se faz nesse sentido é o custo para a avaliação. Minutos pagos, mesmo em rede local, são caros a ponto de afetar o orçamento da avaliação. Em geral, as cadeias de televisão veiculam informações sobre a inscrição (por exemplo, no caso do ENEM), ocorrências durante as provas ou a divulgação dos resultados. Não há divulgação de orientações mais específicas ou espaço para reflexões sobre os achados.

Em relação a como as informações da avaliação chegarão aos *stakeholders*, uma questão – além dos itens verificadores – deve ser posta: quem é responsável pela disseminação da informação, o cliente ou o avaliador externo? Em muitas políticas, a encomenda da avaliação é finda com a entrega das informações, em seus muitos formatos, ao cliente principal, seja ele uma superintendência estadual ou uma diretoria no ministério. Em vários casos, por uma questão contratual, os avaliadores chegam a escrever *release*s, mas estes são encaminhados para a imprensa pelo cliente. A meta-avaliação, portanto, precisa observar quem é responsável pela disseminação dos dados (a logística dos resultados) antes de passar para a análise dos documentos de comunicação utilizados.

Os indicadores U5 e U6 referem-se a aspectos da comunicação da informação em relação aos *stakeholders* e usuários. O U7, último indicador da categoria Utilidade do JCSEE, trata de como as estratégias são usadas e de como é estabelecida a relação avaliador x *stakeholders* de modo a afetar os usos da avaliação. É sobre isso que trata a subseção a seguir.

2.4.2.7 Itens de verificação para o indicador U7 – Impacto da avaliação.

- Mantém contato com o público-alvo.
- Envolve *stakeholders* ao longo da avaliação.
- Incentiva e apóia *stakeholders* na utilização dos achados.
- Demonstra aos *stakeholders* como utilizar os achados em sua prática/trabalho.
- Prevê e endereça usos potenciais dos achados.
- Provê relatos parciais.
- Assegura que os relatórios sejam abertos, francos e concretos.
- Suplementa comunicação escrita com comunicação oral contínua.
- Conduz *workshops* de *feedback* para rever e aplicar os achados.
- Organiza-se de modo a oferecer *follow-up* aos usuários na interpretação e utilização dos achados.

Os itens de verificação do indicador U7 – Impacto da Avaliação apontam para o estabelecimento de uma relação próxima entre avaliador e *stakeholders* como elemento de qualidade em uma experiência avaliativa. Respostas aos itens de verificação desse indicador (que seria melhor denominado “efeito da relação do avaliador com os *stakeholders* no uso da avaliação”) podem ser buscados nos documentos dos programas de avaliação que tratam dos encontros avaliador x *stakeholders*, da análise dos relatórios e outras peças de comunicação, das programações dos encontros e *workshop*s conduzidos pelos *stakeholders* principais ou pelos avaliadores.

Como ocorreu nos seis outros indicadores do JCSEE, há sobreposição dos itens de verificação do U7 com outros discutidos anteriormente. É o caso, por exemplo, dos itens que tratam de relatórios. Esse aspecto já foi abordado antes pelos indicadores U5 e U6, o que mostra que a crítica de Widmer (2005) é pertinente para todos os indicadores da categoria Utilidade. Também da mesma forma que ocorreu com os outros seis indicadores, é possível a utilização dos itens verificadores do U7 para análise de políticas de avaliação. Entretanto, tal transposição deve considerar ajustes quanto à abrangência quando o foco são avaliações em larga escala. A discussão é a mesma apresentada no U1: o que é representativo quando se lida com um grupo de, por exemplo, 2.000 escolas? Autores preocupados com o uso da avaliação argumentam que o envolvimento dos *stakeholders* no processo, em contínuo diálogo, aumenta a probabilidade de uso. Como posto por Abramowicz (1994:81), “a ligação entre avaliadores e quem toma decisões

contribui, decisivamente, para a utilização eficiente dos resultados da avaliação". Como avaliar a proximidade de contato avaliador – *stakeholder* nesses casos?

A relação avaliador x *stakeholder*(s) deve ser verificada ao longo do processo e não só ao final. Um exemplo de estabelecimento da relação avaliador x *stakeholders* no início do processo é o SAEB: na definição das matrizes de avaliação, o INEP envolveu representantes dos estados da Federação. Já para a definição da amostra, os estados foram ouvidos, mas as sugestões não foram incorporadas. Houve reação, especialmente dos Estados do Norte do país, por causa da exclusão das escolas rurais da amostra por dificuldades na logística de aplicação dos testes. De qualquer maneira, houve um esforço de envolvimento dos estados no delineamento da avaliação. O mesmo não ocorreu, por exemplo, com o Sistema Mineiro de Avaliação da Aprendizagem (SIMAVE) em 2000/2001. Ao contrário do SAEB, optou por construir sua matriz de referência apenas com a participação de professores da Educação Superior, sem uma consulta às escolas, o que lhe rendeu alguma crítica. Contudo, não foram encontrados estudos que mostrassem que uma e outra posição tivessem afetado o uso das avaliações.

A relação avaliador x *stakeholder* pode ser observada inclusive como estratégia para que a política de avaliação venha a ser implementada. Esse foi o caso do SINAES. Quando da finalização da proposta, anterior à promulgação da Lei 10.861/2004, representantes do que viria a ser a CONAES visitaram todos os Estados para discuti-la e buscar apoio. O SINAES provê um segundo exemplo de envolvimento avaliador x *stakeholder*, nesse caso para o refinamento do Sistema: a CONAES, em 2008, promoveu um encontro em Brasília com a comunidade acadêmica para discutir o SINAES, no qual estavam presentes os técnicos avaliadores, os coordenadores de cursos e estudiosos contratados pelo INEP.

Para complementar o quadro de itens de verificação para o U7, seria interessante que fosse adicionado o item: "Identifica o repertório para mudança dos *stakeholders*". Como discutido por Weiss (1998), as condições organizacionais precisam mudar – no sentido de remoção de impedimentos, garantia de infra-estrutura e de apoio – para que os resultados sejam utilizados no melhoramento do objeto da avaliação. O *stakeholder* deve ter um nível de autonomia e repertório para que possa lidar com os resultados da avaliação. Nesse panorama, a relação *stakeholder* x avaliador é benéfica e uso-conducente. Caso contrário, o uso é mínimo, independente do esforço do avaliador. Um exemplo disso é o trabalho das Comissões Próprias de Avaliação (CPA) das universidades federais, dentro da proposta do SINAES. A auto-avaliação é conduzida, os resultados são obtidos, mas as universidades não têm autonomia que lhes permita mudar.

Apresentados os sete indicadores da categoria Utilidade, é possível perceber que nenhum deles trata dos usos reais concretizados (ainda que mencionem o atendimento da demanda) ou da percepção dos *stakeholders* sobre a utilidade do processo avaliativo. Como já discutido anteriormente (Subseção 2.4.1), o termo utilidade refere-se a uma percepção do indivíduo, muitas vezes mais atrelada ao atendimento de suas demandas / expectativas que à possibilidade de uso da política. Por essa razão, na adaptação do modelo do JCSEE e do *checklist* para a análise de políticas públicas, foi acrescido o indicador U8 – Percepção da utilidade dos *stakeholders*.

Os setenta itens verificadores de Stufflebeam (1999), neste trabalho reduzidos a sessenta e seis, acrescidos do U8, abordam os elementos que, de um modo ou de outro, favorecem o uso, mas não necessariamente conduzem o meta-avaliador à analise sobre o uso efetivamente concretizado. Nesse sentido, para a análise das contribuições das políticas de avaliação, a categoria Utilidade pode ser aplicada na busca por elementos preditores do uso e deve ser complementada por uma segunda categoria que permita a análise dos usos feitos. A construção dessa segunda categoria – Uso – é feita na próxima subseção a partir de discussão sobre diversos tipos de uso possíveis no contexto da avaliação.

2.4.3 Construção da categoria Uso

Para a construção da categoria Uso, foram considerados os tipos de usos e a discussão dos maus usos da avaliação. O conceito de uso pode ser operacionalizado a partir de diferentes dimensões. Um número considerável de autores, dentre os quais Weiss (1997), tem feito distinção entre uso instrumental e uso conceitual e desdobra essa última categoria em várias outras, como uso político, uso persuasivo, uso simbólico e uso informativo (*enligthment*), que ainda hoje têm aplicação. O uso instrumental diz respeito à utilização direta dos resultados para a tomada de decisões; o conceitual implica mudanças nos pensamentos, atitudes e conhecimentos, sem uma ação imediata. Essa diferenciação enriquece a meta-avaliação de políticas de avaliação e, por essa razão, é detalhada a seguir.

2.4.3.1 Uso Instrumental

O uso instrumental é mais tradicional quando se pensa nos estudos sobre avaliação e nas demandas dos *stakeholders*. Ele está atrelado às ações decorrentes da tomada de decisões informada pelos resultados da avaliação. Para Patton, esse tipo de uso está vinculado à finalidade "julgamento do objeto" (PATTON, 1997:63-85), que Weiss relaciona às decisões no sentido de finalizar uma intervenção, modificá-la ou mantê-la (WEISS, 1998). Pressupõe-se que o uso instrumental se dê quando a avaliação responde as perguntas avaliativas do programa. De acordo

com Weiss, esse tipo de uso acontece sob três condições: 1) se as implicações dos resultados obtidos são relativamente não-controversas, sem que provoquem rupturas na organização ou que afetem interesses conflitantes; 2) se as mudanças pedidas estão dentro do repertório do programa e ocorrem em pequena escala; e 3) se o ambiente no qual o programa acontece é relativamente estável, sem grandes alterações em suas lideranças, orçamento, clientela ou apoio público. Ainda para Weiss, existe uma quarta condição: quando o programa está em crise ou paralisado, sem que ninguém consiga resolver suas questões. Como resposta extrema, recorre-se à avaliação (WEISS, 1998).

As condições apresentadas por Weiss para o uso instrumental raramente são encontradas nos ambientes das políticas de avaliação educacional em larga escala, especialmente as realizadas no Brasil. As implicações dos achados, ainda que não *high stakes* para as fontes, são grandes quando considerados os gestores públicos nas secretarias e ministério de educação ou, mais localmente, nas escolas. Há interesses atingidos quando são pesadas as ideologias dominantes nas faculdades de pedagogia, que repetem, como um mantra, que as avaliações são ferramentas do capital para a exclusão daqueles que já estão à margem. Por exemplo, Frigotto (em entrevista a YAZBECK, 2007:18) reclama que a avaliação no Brasil é um instrumento único para todo o país, apesar de que, por lei, as escolas têm autonomia sobre 25% do seu currículo, devendo adequá-lo para as condições locais ou regionais. A crítica de que a avaliação é única não considera que ela poderia estar voltada para os 75% do currículo comuns a todos os estados e que, localmente, poderia ser complementada por aspectos mais individualizados.

Os resultados, na sua maioria indicando deficiências graves na aquisição de conhecimentos pelos alunos, não são aceitos por serem considerados restritos demais ou por serem fruto de delineamentos a serviço da ordem dominante. Resultados não aceitos não são utilizados. Além disso, as avaliações pedem mudanças que estão fora do repertório das secretarias e das escolas, muito restrito uma vez que essas instâncias têm sido caracterizadas por baixa capacidade institucional. Por fim, o ambiente raramente é considerado estável, com mudanças nas lideranças políticas a cada quatro anos e com redefinições de orçamento para diferentes programas que afetam o programa ou sistema sob avaliação e a política de avaliação em si. A pensar nessa definição de Weiss, apenas teria uso instrumental a avaliação resposta a uma crise.

Uma segunda reflexão se faz necessária. Um programa de avaliação é uma política pública, como definido na anteriormente neste Marco Teórico. Os estudiosos sobre políticas públicas demonstram que a tomada de decisões nem sempre é racional (ou não é racional). Há uma série de

condições que a afetam, como pode ser visto em Abramowicz (1994), para quem administradores e políticos não valorizam os resultados da avaliação “na medida em que eles são componentes de um determinado sistema político com seus próprios valores e estão envolvidos com preocupações que extrapolam aquelas sobre a eficiência de um programa". Essa posição é defendida também por Miriam Warde que, em entrevista a Yazbeck (2007:18), relata que “as políticas têm sido traçadas independentemente das avaliações”.

Por outro lado, como posto por Creso Franco ao refletir sobre pesquisa, os intelectuais brasileiros, especialmente desde a década de 80, passaram a ocupar cargos nos sistemas educacionais e, ao fazê-lo, trouxeram e trazem suas “bagagens intelectuais e de pesquisa para os postos em que atuam”. O complicado é avaliar os “diversos padrões de relação estabelecidos entre os campos acadêmicos e políticos e, consequentemente, entre pesquisa e construção de políticas educacionais” (em entrevista a YAZBECK, 2007:16). A relação decisão x uso da informação não é tão direta quanto o uso instrumental pressupõe que seja. Na mesma investigação de Yazbeck, diz Zákia Souza sobre o uso da pesquisa:

> Nos limites das considerações aqui registradas quero destacar apenas dois pontos[...]: a) a literatura sobre a utilização das pesquisas na formulação e reformulação das políticas públicas tem evidenciado que esse processo não se dá de modo linear, demandando, com raras exceções, um tempo para que os gestores incorporem eventuais contribuições das pesquisas no delineamento das políticas; e b) a incorporação das contribuições das pesquisas muitas vezes se evidencia tanto no plano da legislação como nos planos e programas governamentais. No entanto, há que se observar como se deu tal incorporação, ou seja, como essas contribuições foram interpretadas à luz do programa governamental mais amplo, bem como que condições concretas foram viabilizadas para tornar realidade uma dada proposta que emerge das investigações. (Zákia Souza em entrevista a YAZBECK, 2007:16).

O mesmo pode ser dito sobre o uso de informações derivadas dos delineamentos das políticas de avaliação.

Estabelecer a relação (direta) entre informação e decisão, no caso de formulação das políticas, não é, portanto, tarefa simples. De todo modo, não deveria ser a tarefa única também. Como posto por Ferrer (1997), deter-se apenas aos usos instrumentais é simplificar demais um processo complexo: é importante considerar também os usos conceituais.

2.4.3.2 Uso Conceitual

A dificuldade na busca do uso instrumental permanece no levantamento do uso conceitual, que se dá quando os usuários não têm condições de utilizar instrumentalmente os achados, mas tais

resultados mudam sua percepção sobre o programa e seus efeitos. Eles obtêm, assim, novas idéias e *insights*; podem aprender sobre os pontos fortes e fracos e possíveis linhas de ação. Quando as condições de contexto tornam-se mais favoráveis, é possível que esse conhecimento e esse novo entendimento transformem-se em ações e o uso se dê de modo instrumental (WEISS, 1998). Um exemplo disso pode ser o uso que algumas unidades das universidades federais fazem a partir da auto-avaliação, componente do SINAES. Ainda que, por falta de autonomia administrava, não possam mudar, formam uma idéia sobre o que precisa ser mudado (RIBEIRO, 2009). A categoria uso conceitual aproxima a utilização da avaliação do conceito de influência, apresentado na subseção anterior.

Além da muito falada ausência de autonomia no nível micro (universidade e escolas), a falta de uso conceitual, assim como no uso instrumental, pode ser associada à carência de capacidade instalada nas instâncias governamentais envolvidas com a educação. No caso dos resultados das avaliações em larga escala, o que se vê é uma distância tão grande entre o nível ideal e o real que as escolas e as secretarias ficam paralisadas. Seu repertório é insuficiente para propor mudanças e, embora reconheçam o nível precário da educação pública em seus sistemas e micro-sistemas, não têm a menor concepção de como devem contribuir para mudá-lo.

Em estudos sobre o uso conceitual, um outro problema se coloca: para seu levantamento é essencial que se busque o relato das fontes, mas as respostas obtidas podem ser fruto de um comportamento "conformado". No caso das escolas, há uma incorporação do jargão de avaliação no seu discurso corriqueiro, sem que isso implique mudança de percepção ou de atitude (DANTAS, 2005). Ao responder questionários sobre os usos potenciais da avaliação em seus ambientes de trabalho, vários usuários optam por responder "a resposta certa", em lugar de apresentar seu real posicionamento.

A ritualização poderia ser classificada como uma estratégia de sobrevivência da "burocracia ao nível da rua" (LIPSKY, 1980) na escola pública que, a cada mudança de governo, vê-se às voltas com a implementação de novos programas e novos projetos, sem necessariamente terem relação com suas crenças e com suas práticas (DANTAS, 2005). Em lugar de contestar essas imposições dos órgãos centrais, as escolas implementam tais políticas de maneira ritual ou conformada, como explicado por Meyer e Rowan (1991), citados por Libório e Costa (2004:698):

> [...] as organizações incorporam os procedimentos definidos por forças institucionais exteriores, tais como a opinião pública, os sistemas educativos, as leis, os tribunais, as profissões, as ideologias, as tecnologias e as estruturas reguladoras. Ao incorporarem os procedimentos provenientes dos vários

> meios institucionais, as organizações aumentam a sua capacidade de sobrevivência e a sua legitimidade, independentemente da eficácia dos procedimentos adoptados. Assim a legitimidade organizacional depende não da eficácia dos procedimentos, mas da conformidade com os meios institucionais. Nesta perspectiva, as técnicas, as políticas e programas que se institucionalizam nas organizações funcionam como "mitos racionais", adoptados cerimonialmente, o que permite a conformidade com as regras culturais dominantes. Deste modo, as organizações garantem a sua legitimidade e aceitação social, ou seja, a sua sobrevivência.

Seria muito importante que, no Brasil, estudos aprofundados fossem conduzidos sobre o uso conceitual, não só junto aos usuários-fim, mas também aos formuladores das políticas educacionais, *stakeholders*-chave que, em diversas ocasiões, estão apenas praticando a pseudo-avaliação (STUFFLEBEAM, 1974; VIANNA, 1998; RAVELA *et alii*, 2008). A ritualização pode ocorrer também com os formuladores, quando propõem avaliação apenas para responder à pressão social por maior transparência, por exemplo, sem uma função real.

Há variações dos usos conceituais. Dentre elas, Weiss refere-se ao uso persuasório, simbólico, que se dá quando os resultados permitem apoio a uma determinada posição que o *stakeholder* já detém sobre mudanças que precisam de implementação no programa. Frequentemente o gestor do programa e seus implementadores conhecem os erros e o que deve ser feito para corrigi-los. Eles utilizam a avaliação para legitimar sua posição e conseguir adeptos. A avaliação torna-se, assim, um instrumento de persuasão (WEISS, 1998). É interessante perceber que, no Brasil, várias políticas são implementadas a partir de indicadores que não estão nem remotamente relacionados com seu objeto. Seria o caso, por exemplo, da utilização do IDH para justificar ações da educação especial ou ainda dos resultados do SAEB na persuasão de comunidades para novas políticas de EJA.

Por fim, dentre os usos conceituais, Weiss propõe ainda o uso informativo ("*enlightenment" kind of use*), que se dá "extra-muros", para além do programa sob avaliação. O uso informativo é o que aproxima a avaliação da pesquisa científica, por contribuir para o corpo de conhecimento existente. É o que ocorre com meta-análise, quando uma experiência de avaliação é analisada em conjunto com outras similares e fornece um panorama geral de um determinado objeto. Muitas vezes, a meta-análise é feita por revisão qualitativa. Os relatos das avaliações acabam disseminados pelo campo e podem influenciar as redes de avaliadores e estudiosos de políticas públicas, alterar paradigmas de políticas, mudar a agenda política, e/ou afetar crenças de determinados grupos em instituições. O conhecimento é uma das forças na formulação das políticas e dos programas. Quando a avaliação contribui para o acúmulo de conhecimento, pode

afetar movimentos teóricos que, eventualmente, resultarão em ações. De acordo com Weiss (1998), já houve uma série de estudos sobre esse tipo de uso que mostram que tal categoria não deve ser negligenciada. Há inclusive uma linha de pensamento de avaliação voltada para a teoria, que incentiva o uso informativo na essência.

Um dos exemplos de uso informativo se dá quando os avaliadores consultam experiências anteriores para levantar possíveis "efeitos colaterais" para o novo desenho a propor. Por exemplo, a determinação da matriz avaliativa educacional em nível nacional pode trazer enrijecimento de currículos e, em muitos casos, isso é considerado efeito colateral. Concorda-se com Sousa (2003) quando diz que, na verdade, esse tipo de efeito é intrínseco ao delineamento e, nesse caso, é possível prevê-lo se se acessam experiências avaliativas anteriores.

Usos conceituais e instrumentais não originalmente pretendidos muitas vezes estão distantes da finalidade avaliativa, mas não necessariamente são maus usos ou abusos, apenas usos não previstos. Patton (1988b *apud* SHULHA; COUSINS: 1997:202-204), em um outro arcabouço teórico, opõe utilização a não utilização e não má utilização a má utilização, sendo a intencionalidade a variável que distingue as duas últimas. Propõe-se aqui que o uso ritual seja tratado como uma não utilização (e não uma má utilização).

O primeiro passo na promoção de um diálogo com os *stakeholders* sobre a má utilização x não má utilização pode ser encorajar o relato de pontos não antevistos de decisão e conseqüências não previstas nos planos de ação quando se conduzem estudos sobre uso. Nessa linha, para Bamberger, Ruth e Madry (2006), o mau uso pode ser intencional, mas pode também ser decorrente da incapacidade de interpretação de dados e resultados por parte dos usuários e *stakeholders*, ou ainda pela má comunicação dos mesmos. Moura Castro, em artigo de 2001 escrito para o INEP (CASTRO, 2001), mostra, por exemplo, o desserviço prestado pela imprensa ao divulgar resultados do Provão, o que pode vir a causar interpretações errôneas com conseqüências nefastas, especialmente quando se considera a escolha de uma determinada universidade pela pontuação obtida em uma avaliação.

A leitura de muitos dos estudos sobre os usos e a utilidade da avaliação mostra que não há um modelo único para abordá-los. Como acontece no geral no campo da avaliação, a área de conhecimento do pesquisador (bem como sua experiência profissional) favorece uma ou outra abordagem. Nesse trabalho, optou-se por criar uma categoria Uso em duas dimensões: uso

instrumental e uso conceitual, a partir da discussão apresentada anteriormente. A Ilustração 06 a seguir sintetiza os indicadores de Uso utilizados na presente pesquisa.

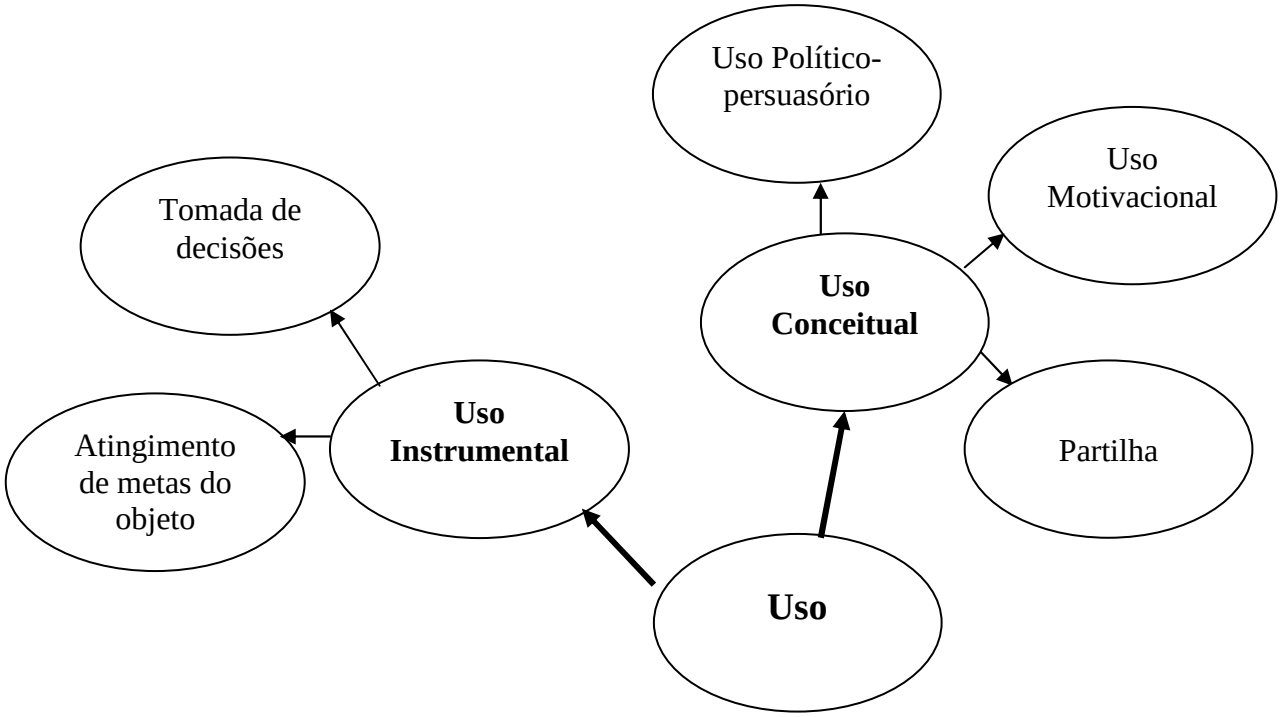


Ilustração 6: Síntese da categoria Uso como utilizada no presente trabalho de pesquisa.

Com a síntese da categoria Uso, encerra-se aqui a apresentação da Fundamentação Teórica utilizada para o estudo em tela. O foco da investigação é a política de avaliação de aprendizagem (AA), formulada em 1999 pela Secretaria da Educação do Estado da Bahia, como o elemento de avaliação em um programa de reforma educacional. A AA foi implementada pela Universidade Federal da Bahia, com aplicações de provas a partir de 2001. Os *stakeholders* principais foram as escolas públicas estaduais e municipais localizadas nas áreas urbanas de municípios envolvidos na referida reforma, promovida até 2004. A descrição da política foco abre, a seguir, o capítulo 3. Metodologia para que, na seqüência, sejam apresentados a lógica da pesquisa e os quadros de análise e operacional empregados.

3. Metodologia

3.1 A política foco da presente investigação

Como foco do presente estudo, que busca compreender as contribuições das políticas de avaliação, escolheu-se o primeiro ciclo da política *Avaliação de Aprendizagem (2001-2004)*, implementada sob a coordenação do Centro de Estudos Interdisciplinares para o Setor Público (ISP) da Universidade Federal da Bahia (UFBA), para a Secretaria da Educação do Estado da Bahia (SEC), junto a um universo de aproximadamente 2.850 escolas públicas urbanas, estaduais e municipais, localizadas em 299 municípios do Estado, na última unidade de 2004[45]. A parceria da SEC com a UFBA foi estabelecida por meio de convênio (444/99) e contou com a interveniência da Fundação de Apoio à Pesquisa e à Extensão (FAPEX). Todas as atividades foram desenvolvidas sob o nome Projeto de Avaliação Externa / Agência de Avaliação UFBA-ISP/FAPEX, referido, no presente texto, por Projeto de Avaliação Externa. Para desenvolvimento da política, foi contratada uma equipe externa à UFBA e à SEC, aqui referida como equipe central da avaliação.

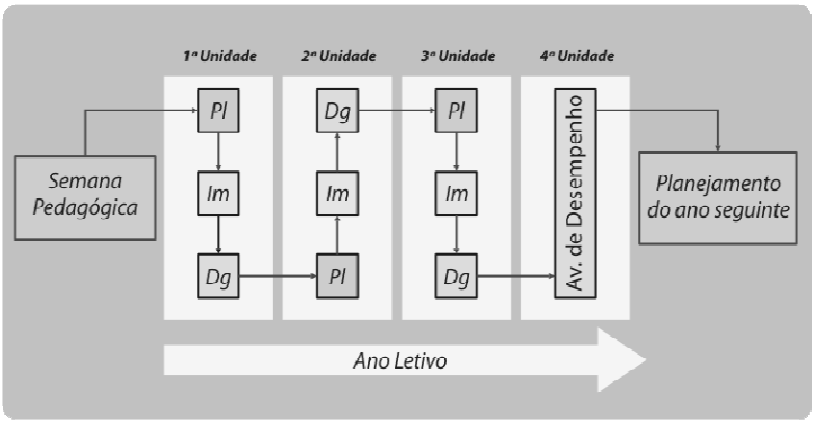
A escolha da política foi feita porque, discrepante em relação a outras experiências de avaliação no Brasil, a *Avaliação de Aprendizagem* (AA) teve um caráter menos regulatório que as demais políticas implementadas à época, vez que seu objetivo foi instrumentalizar as unidades escolares públicas com ferramentas diagnósticas a cada 200 horas letivas. Pretendia-se, com essa abordagem, que docentes e gestores escolares pudessem flagrar problemas na aquisição de competências e habilidades pelos alunos, durante o ano letivo, a tempo de serem colocadas em prática ações de remediação. Essas ações, em último caso, resultariam em uma menor taxa de reprovação e no oferecimento de um melhor serviço educacional.

Utilizando-se os adjetivos da avaliação brevemente apresentados na Subseção 2.2.2, pode-se dizer da AA que foi uma experiência avaliativa longitudinal (ainda que não tivesse voltada para a comparabilidade ao longo do tempo), formativa, de caráter diagnóstico e *low stakes*. Sua especificidade está em, não sendo uma avaliação da escola e certamente não devendo substituir a avaliação na escola, poder ser considerada "uma avaliação para a escola". Como política, foi

45 Universo expandido gradualmente; em 2001, eram 126 os municípios parceiros do Estado no Programa Educar para Vencer. O estabelecimento da parceria entre o Estado e o Município envolvia uma negociação que incluía, dentre outros aspectos, a implementação de todos os projetos do Educar e o compromisso do município no apoio às ações.

formulada centralmente para uma implementação local, com expectativa de impacto também localizado.

A AA foi uma das duas vertentes do Projeto 06 – Avaliação Externa, integrante do Programa Educar para Vencer, apresentado brevemente a seguir na subseção Contexto. A segunda vertente do Projeto chamou-se Avaliação de Desempenho (AD), formulada e implementada em uma adequação do *National Assessment of Educational Progress (*NAEP) americano e do SAEB, cujas fontes foram alunos da 4ª e 8ª série do Ensino Fundamental e, em 2004, também do 3º ano do Ensino Médio. Os recursos para as ações foram provenientes do Governo do Estado e do Banco Mundial. Juntas, as duas vertentes de trabalho deveriam resultar na formação de uma cultura de avaliação em larga escala na Bahia. A Ilustração 7 a seguir traz a representação de como as duas avaliações deveriam ser percebidas pelas escolas.



Legenda: pl = planejamento; im = implementação; e dg = diagnóstico fornecido pela AA.
Fonte: Projeto de Avaliação Externa

Ilustração 7: Lógica do relacionamento entre a Avaliação de Aprendizagem e a Avaliação de Desempenho no contexto do Educar para Vencer.

Durante a semana pedagógica, idealmente, as escolas teriam acesso a seus resultados da AD, realizada no final do ano anterior[46]. Com esses dados, poderiam planejar seu plano de cursos para o ano vigente prevendo o combate aos pontos fracos apontados. Durante o ano, a cada unidade letiva (200 horas), a escola teria o apoio da AA, com provas padronizadas e resultado imediato, já que ela mesma corrigiria os testes. Esse diagnóstico favoreceria um replanejamento ao longo do ano letivo. O esforço seria avaliado pela AD, novamente, ao final do ano. Da AD foram utilizadas as médias, por escola, do desempenho de seus alunos de 4ª série em Português e

46 Na implementação da AD, os resultados – em boletins individualizados – foram entregues às escolas em abril / maio do ano seguinte ao da aplicação de provas. Dessa maneira, não foi possível utilizar seus dados durante a Semana Pedagógica, usualmente realizada em janeiro/fevereiro. Entretanto, esses dados poderiam ser utilizados quando das discussões de replanejamento a partir do 1º diagnóstico da AA.

Matemática em 2004, como um elemento de análise para a própria política de AA, a ser detalhado na Subseção 3.1.2 adiante.

O estudo sobre a AA, cuja interação da equipe central de avaliação com os *stakeholders* principais deveria ser intensa, pode colaborar para uma maior compreensão das contribuições de uma política de avaliação, pela análise dos usos feitos (não só dos resultados), bem como da sua utilidade para as escolas públicas. Entretanto, para apresentá-la propriamente ao leitor, é fundamental que a política esteja inserida no contexto político maior, de modo a favorecer as discussões sobre os usos e utilidade relatados pelas escolas públicas. É desse contexto que trata a subseção a seguir.

3.1.1 Contexto

Em 1999, o Governo do Estado da Bahia, através da SEC, lançou um programa de reforma chamado Educar para Vencer, com foco no aluno e "eixo da mudança" na escola. Com uma taxa de atendimento aos jovens de 07 a 14 anos superior a 95%, o Governo buscou concentrar esforços na regularização do fluxo de alunos (70% de defasagem na 5ª a 8ª séries), no combate às taxas altas de abandono e repetência, e na "profissionalização" da gestão escolar. Dirigentes da rede estadual eram (e continuam a ser) escolhidos no quadro de professores concursados; entretanto, não havia qualquer preparação para que esses professores assumissem cargos de direção, cuja demanda se tornava cada vez mais complexa em um ambiente que visava à autonomia escolar. Além dessas questões, o discurso oficial reconhecia que o fato de o aluno estar na escola não significava necessariamente que estivesse recebendo uma educação de qualidade[47], reforçado por resultados no SAEB não satisfatórios.

O objetivo máximo do programa de reforma – sua finalidade, de acordo com conceituação já feita (Subseção 2.4.1) - foi melhorar a qualidade da educação pública no Estado, "expressa através do sucesso escolar dos alunos" (Manual do PDE, 2001). Para alcançar seu objetivo, o Programa foi desenhado com base na articulação de seis projetos prioritários e três ações complementares, fundamentados nos princípios de autonomia, eqüidade, ação local e qualidade. No contexto da presente pesquisa, entende-se cada um desses projetos e das suas linhas de ação como política pública (*policy*), conforme conceitos de Dye (1995) e Souza (2002), discutidos anteriormente na

47 Ver *folder* " Educar para Vencer, o ensino público do novo século", primeiro documento utilizado pela SEC para divulgação do Programa de Governo (BAHIA, 1999).

Subseção 2.1.1. As ações foram voltadas para o Ensino Fundamental, sem esquecer as questões de municipalização das escolas de 1ª a 4ª série e a expansão do Ensino Médio.

As equipes contratadas para implementar os projetos prioritários, especialmente os de gestão, certificação e avaliação, foram compostas por profissionais sem vínculo anterior com o serviço público. As equipes dos dois projetos de regularização foram mistas, compostas também por profissionais da própria SEC. O quadro a seguir apresenta os objetivos dos seis projetos prioritários, divulgados através de *folder*s promocionais publicados pela SEC, em 2000, e de manuais e documentos produzidos pelas diversas equipes dos seis projetos.

Programa Educar para Vencer	**Finalidade: Melhorar a qualidade da educação pública do Estado da Bahia**
Projeto	**Objetivos específicos**
Regularização do Fluxo Escolar 1ª a 4ª série	1. Regularizar o fluxo escolar dos alunos das séries iniciais do Ensino Fundamental da rede pública estadual e municipal. 2. Trabalhar, junto aos alunos com defasagem série x idade, um currículo especial que contemple conteúdos básicos para as séries iniciais, permitindo-lhes avançar para a 5ª série, quando comprovado o alcance dos objetivos do curso. 3. Prover, para os alunos que não têm o domínio da leitura e escrita, uma fase de alfabetização.
Regularização do Fluxo Escolar 5ª a 8ª série	1. Regularizar o fluxo escolar dos alunos matriculados nos dois segmentos, A (5ª série) e B (6ª e 7ª séries). 2. Resgatar a auto-estima dos alunos defasados, ensiná-los a aprender a aprender, com ensino de qualidade. 3. Desenvolver competências e habilidades para continuidade de estudos nas séries compatíveis com a sua idade. 4. Prover materiais a docentes e discentes, capacitação aos docentes e acompanhamento sistemático do processo.
Fortalecimento da Gestão Escolar	1. Reordenar, estruturar, instrumentalizar e capacitar dirigentes de escolas públicas estaduais e municipais para operarem de forma autônoma, contribuindo para o aprimoramento constante da qualidade do ensino e para a racionalização do número de escolas. 2. Assegurar a melhoria constante da qualidade de ensino, através da implementação de um padrão de gestão profissional.
Capacitação Gerencial das Unidades Municipais da Educação	1. Fortalecer a gestão municipal, de forma a compatibilizar e otimizar as ações do poder público e a promover a operação de escolas autônomas.
Certificação dos Profissionais da Educação	1. Propiciar à SEC a seleção de profissionais capacitados e a manutenção no sistema de um quadro de profissionais qualificados. 2. Assegurar não somente que os profissionais dominem as competências necessárias ao exercício dos cargos que ocupam, como também que se mantenham atualizados em relação à contínua evolução que se processa no seu campo de conhecimento (HASHIMOTO, 2003).
Avaliação Externa	1. Fornecer às escolas ferramentas diagnósticas e informações que possibilitem a identificação de problemas e a conseqüente melhoria do ensino oferecido. 2. Fornecer a cada escola os resultados alcançados por seus alunos em relação àquilo que deveriam estar aprendendo.

Fonte: DANTAS, 2005

Quadro 6: Síntese dos objetivos específicos dos seis projetos prioritários do Programa Educar para Vencer, como divulgado em *folder*s e materiais promocionais.

Quando do lançamento da reforma, a proposta do Estado previa grande modificação, inclusive em sua base legal, de modo a favorecer a autonomia escolar e a articulação estado – município, dentre outros aspectos[48]. Entretanto, na SEC e no período 1999-2002, as mudanças estiveram voltadas para a reorganização do espaço físico, a contratação da equipe de implementação da política, e a adequação da infra-estrutura, além de um incentivo em forma de gratificação para gestores escolares estaduais a partir do cumprimento de algumas metas. Portarias estaduais específicas e editais foram publicados em relação ao processo de certificação e à atuação dos projetos de fluxo (REIS, 2003), mas não houve uma mudança significativa nas estruturas burocráticas existentes. À época, além do deslocamento dos recursos para os projetos prioritários, a ênfase nos aspectos de fortalecimento da gestão, chave do Programa Educar para Vencer, desagradou os setores da burocracia média voltados para as questões pedagógicas, tanto no órgão central quanto nas Diretorias Regionais (DIREC). A burocracia média resistiu de maneira clara, como pode ser visto pelo lançamento, em 2000, da proposta de Educação Básica da Superintendência de Ensino - SUPEN (Construindo a Escola Terra Bahia), ao esclarecer o conceito de projeto político-pedagógico:

> [...] Em muitas escolas, nas primeiras elaborações, houve certa confusão entre projeto político-pedagógico e plano de desenvolvimento da escola mais conhecido por PDE. Tratava-se de uma idéia tecnicista que valorizava apenas o preenchimento de formulários e aplicação de fórmulas para o cálculo dos índices de aprovação, reprovação e evasão, mas não se comprometia com a análise qualitativa desses dados e nem tampouco com a efetiva alteração da realidade. (SEC/SUPEN, 2000:19)

O texto traz uma crítica explícita ao PDE (Plano de Desenvolvimento da Escola), um dos principais instrumentos de trabalho da equipe de Fortalecimento da Gestão Escolar junto aos dirigentes. Para este grupo, o PDE foi definido como "documento que reflete intenções, objetivos, metas e ações, visando transformar a realidade da Unidade Escolar existente na desejada" (Manual do PDE, 2001) e ferramenta para possibilitar o repasse de recursos às escolas estaduais (as municipais não contaram com isso, o que, de início, já desequilibrava a implementação do Programa).

No desenho do Educar para Vencer, não havia uma linha de ação para suporte técnico-pedagógico às escolas em relação a seus cursos regulares, vez que os projetos de Gestão estavam direcionados para o suporte administrativo gerencial e os de Regularização tratavam de uma clientela especial, defasada em mais de dois anos. As escolas supostamente incluiriam as demandas de capacitação e reforço para as ações pedagógicas em seus PDEs, no caso de recebimento de recursos, e/ou

48 Um exemplo de re-estruturação pode ser visto no Estado do Maranhão, no mesmo período.

contariam com o apoio dos seus órgãos centrais. Em relação à SEC, a resistência aos projetos do Educar foi um empecilho para que o apoio às escolas fosse eficaz, potencializado pela escassez de recursos nos orçamentos dos diversos departamentos e diretorias, vez que a maior monta esteve atrelada aos projetos prioritários. Além disso, a busca pela qualidade deveria ser pautada por atingimento de certos padrões mínimos de funcionamento da escola, trabalhados em 1999-2000, mas que não chegaram a ser definidos por portaria ou implementados pela SEC.

A não adoção dos padrões mínimos não permitiu alteração do quadro de problemas de infra-estrutura e de recursos humanos que caracterizava a escola pública, com conseqüências para a implementação de ações previstas nos seus planos de desenvolvimento. Até o final de 2003, o Educar para Vencer não havia alterado a organização escolar, a não ser pela capacitação dos dirigentes, pela compra de equipamentos e materiais, por um lado, e pela tentativa, bem sucedida na capital com as escolas estaduais, de introdução de mérito como critério para escolha de dirigentes (através da Certificação) até meados de 2003. A rede municipal, de maneira geral, não aderiu a esse princípio e a escolha dos dirigentes continuou ligada à definição político-partidária.

De qualquer maneira, com ou sem recursos para o PDE, o processo de autonomia escolar não foi completo, a não ser em um piloto realizado em duzentas escolas da rede estadual (REIS, 2003). Mesmo assim e sem padrões mínimos assegurados, as escolas estaduais passaram a sofrer não só com as pressões dos projetos prioritários, como também outras demandas do próprio órgão central e das DIREC, além de estarem expostas às políticas e solicitações locais.

Quanto à relação Estado – Município, nem sempre foi possível uma articulação entre a SEC e as secretarias municipais, especialmente na esfera técnica. As escolas municipais, embora livres das demandas das representações regionais da SEC, estiveram ligadas às coordenações pedagógicas de suas secretarias de educação, muitas vezes implementando, em paralelo, outros programas percebidos como conflitantes com os procedimentos adotados pelo Educar para Vencer (DANTAS, 2005). Somando-se às dificuldades de articulação do programa prioritário do governo com a burocracia instalada nas estruturas existentes no Estado e nos municípios, estiveram aquelas voltadas para a articulação entre os seis projetos prioritários que, à medida que foram implementados, distanciaram-se da proposta original.

Para agravar esse distanciamento, ainda em 2000, um grande projeto com financiamento do Banco Mundial, proposto pela gestão anterior de governo (Projeto de Educação do Estado da Bahia - Projeto Bahia), foi finalmente aprovado. Diante da escassez de recursos do Estado, houve

uma tentativa de articular as ações do Educar para Vencer com aquelas financiadas pelo Banco Mundial, ainda que o primeiro estivesse voltado para o Ensino Fundamental e questões de gestão e o segundo tivesse também interesse no Ensino Médio, no re-ordenamento da rede e em construção de edificações. A priorização dos seis projetos do Educar para Vencer passou então a ser relativa, com certa dispersão do foco original do Programa. Em *folder* de divulgação do Projeto Bahia em 2000, previa-se que haveria "uma avaliação bianual da rede pública de ensino realizada em processo gradual de atendimento às unidades escolares localizadas nos 100 municípios do Programa Faz Cidadão e nos municípios sede de DIREC" e "avaliação contínua realizada em todas as séries do ensino fundamental, nas escolas localizadas nos municípios do Programa Educar para Vencer". O *folder* informava ainda que, para a segunda fase do Projeto (2003-2004), deveria haver "aumento de proficiência para o Ensino Médio e o Fundamental nas matérias de Português e Matemática em 5%". Não havia indicação de como seria obtido esse percentual.

Nesse cenário, se deu a implementação do programa de reforma, com previsão de expansão progressiva, por meio de parceria Estado x prefeituras, definida a partir de assinatura de um termo de adesão, publicado em Diário Oficial. Em 1999, 45 municípios firmaram essa parceria; a expectativa do Governo era de abranger, até 2003, os 417 municípios da Bahia. Em realidade, ao final de 2003, o Educar para Vencer tinha sido implementado em 299 municípios, aproximadamente, já que em alguns não foi possível implantar os seis projetos em conjunto, como no caso da capital, Salvador.

Entre 1999-2002, o sistema estadual de ensino, além das escolas, era formado pelo órgão central (SEC), com quatro grandes superintendências – Superintendência de Ensino (SUPEN), Superintendência de Políticas e Diretrizes Educacionais (SPDE), Superintendência de Articulação Municipal (SUPAM), e Superintendência da Gestão Escolar (SUPEC) -, uma diretoria geral e uma Coordenação de Projetos Especiais (COPE); 31 diretorias regionais (DIREC); representações nos municípios (coordenadores estaduais); e com o Conselho Estadual de Educação (CEE). A mudança estrutural ocorreu em 2003, com a mudança de governo. As superintendências sofreram alteração e o Projeto de Avaliação Externa, antes ligado à SPDE, ficou ligado à nova Superintendência de Avaliação e Acompanhamento da Educação Básica (SUPAV). Os dois projetos de regularização de fluxo foram unificados, assim como os dois de fortalecimento da gestão (escolar e municipal).

Embora o partido do governo tivesse se mantido no poder, o período 2003-2006 foi aberto por um secretário de educação que se afastou logo no início do mandato, foi substituído interinamente pelo seu chefe de gabinete até que, em abril de 2003, um novo secretário assumiu a SEC. Ainda que tivesse se comprometido a manter o Programa Educar para Vencer, o novo secretário voltou o foco de sua gestão para a formação e a capacitação docentes e se afastou dos princípios do programa de reforma, especialmente quanto à gestão autônoma da escola. O programa de reforma perde aí seu caráter prioritário e passa a ser visto como mais um conjunto de projetos dentre os da SEC.

Em relação à avaliação, o novo secretário demonstrou posição contrária à aplicação de provas padronizadas. Em outubro de 2004, foi findo o convênio da SEC com a UFBA, sem o estabelecimento de nova parceria. Uma equipe de transição foi mantida pela SEC para as ações de AD já previstas para o final de 2004 e para a implementação da AA em 2005. Ao final desse período, o Projeto de Avaliação Externa foi descontinuado. As discussões sobre instabilidade no contexto político feitas por Weiss (1997, 1999) e apresentadas anteriormente na Fundamentação Teórica podem ser aplicadas, sem prejuízo, ao contexto Bahia.

É nesse contexto que, de 2000 a 2004, foi implementada a AA, foco do presente estudo, detalhado a seguir.

3.1.2 A política de Avaliação da Aprendizagem do Programa Educar para Vencer

De acordo com o *folder* de divulgação do Projeto de Avaliação Externa, distribuído em 2003, a AA tinha por objetivos:

> - Fortalecer nas escolas o hábito de desenvolver e cumprir um plano de ensino dentro de prazos pré-estabelecidos.
> - Possibilitar aos professores o diagnóstico dos sucessos e das dificuldades de seus alunos em relação a um elenco de competências e habilidades mínimas definidas para o Estado.
> - Ajudar os professores a reformularem, quando necessário, seu plano de ensino para melhor atender seus alunos, contribuindo para evitar a repetência. (*Folder* de divulgação. Educar para Vencer. Projeto de Avaliação Externa. SEC/UFBA 2003).

De 2001 a 2004, três vezes ao ano, testes de Língua Portuguesa e Matemática foram administrados para alunos de 1ª a 4ª séries e do Ciclo Básico de Aprendizagem I (CBAI). Com isso, as aplicações envolveram todos os alunos (ingressantes a concluintes) do Fundamental Menor, na expectativa de, ao corrigir os problemas logo no início, impactar positivamente o fluxo educacional. Estas provas verificavam o domínio de competências e habilidades associadas às

unidades letivas, de forma não cumulativa. Os alunos responderam as provas com marcações no próprio caderno[49] e suas respostas foram transferidas para um quadro-diagnóstico, que possibilitaria ao professor diagnosticar a proficiência daquela turma com referência a um conjunto de competências e habilidades passível de medição através de itens objetivos.

Nesse cenário, os próprios professores aplicaram e corrigiram as provas, sob coordenação do coordenador pedagógico ou, na sua ausência, do próprio dirigente escolar[50]. Os procedimentos e as orientações foram encaminhados às escolas em manuais e materiais explicativos, lembrando-as de que quebras no padrão de aplicação dos testes e da sua correção teriam implicações sérias nas informações resultantes, podendo comprometer qualquer análise que delas viesse a ser feita. Uma discussão dos resultados obtidos deveria ser feita em reunião da coordenação/direção com os vários professores regentes das turmas avaliadas, de modo a permitir à escola replanejar suas ações para melhor atender a seus alunos. Obviamente, os objetivos da AA só seriam minimamente atingidos quando, após a obtenção do diagnóstico, as escolas analisassem seus resultados e agissem a partir dessa análise. O acompanhamento de toda a ação foi feito através do *Relatório do Diretor* (RD), documento encaminhado pelas escolas à equipe central da avaliação a cada aplicação de provas. Um estudo em uma amostra controlada foi conduzido, a cada aplicação, para permitir um diagnóstico geral das escolas envolvidas pelo sistema de avaliação[51], de maneira a informar as decisões sobre as políticas a serem criadas, além de verificar as tendências apontadas pelas escolas nos RDs.

Todos os testes utilizados foram construídos após análise de itens a partir de aplicações piloto. Em diversas ocasiões, estudos paralelos foram realizados na busca de elementos para refinamento da logística de aplicação e da comunicação com as escolas e para compreensão dos movimentos que as escolas estavam conduzindo na busca da solução dos problemas diagnosticados. As aplicações piloto e as amostras de monitoramento tiveram seus resultados sistematizados em Relatórios Psicométricos e Relatórios da Coordenação de Aplicação de Instrumentos. Esses relatórios, juntos às bases de dados oriundas do processamento dos RDs, formaram o lastro para, em 2004, a elaboração de relatórios síntese que retornaram às secretarias municipais de educação e à SEC,

---

49 A exceção da 4ª série, cujos testes passaram a ser acompanhados por gabaritos, em um atendimento à demanda apresentada pelas escolas, que queriam que os alunos tivessem experiência com o formato utilizado pela AD.
50 De acordo com informações das escolas na 1ª unidade de 2004, 60,6% das 797 escolas estaduais e 28,5% das 485 escolas municipais não contavam com esse profissional, dentre as 1.349 escolas que encaminharam o RD à equipe central da avaliação a tempo de constarem do Relatório Síntese da AA 2004 – 1ª unidade. Fonte: Projeto de Avaliação Externa, Relatório Síntese da AA 2004 – 1ª unidade.
51 O delineamento da amostra, sob responsabilidade de Carlos Henrique Nunes, era representativo ao nível das DIRECs envolvidas na AA.

com os resultados das amostras e as principais ocorrências detectadas durante a aplicação das provas e discussão dos resultados nas escolas.

Desde o início, portanto, a escola assumiu um papel central na implementação da AA. Não houve nenhum ato de avaliação definido através de decreto, municipal ou estadual, ao longo dos quatro anos do primeiro ciclo da AA. Todo o processo foi implementado pelo envio de material de aplicação, de correção e dos manuais explicativos do Projeto de Avaliação Externa para as escolas, para as DIREC, para a Superintendência de Ensino (no órgão central), e para as coordenações pedagógicas das secretarias municipais envolvidas. Essas deveriam apoiar as unidades escolares nas soluções propostas, como parte na parceria estabelecida com a SEC, cujo termo era publicado em Diário Oficial. Em nenhum momento os resultados da AA deveriam ser usados para punir ou premiar escolas ou profissionais. Vídeos didáticos, acompanhados por manuais, foram elaborados e encaminhados às escolas, contendo várias abordagens pedagógicas sobre os descritores para os quais os alunos da amostra tinham demonstrado desempenho mais baixo, como material de apoio à remediação. Os materiais de remediação foram desenvolvidos tanto para utilização com os alunos, em sala de aula, como material de capacitação docente.

A expansão da AA, como as demais políticas do Educar para Vencer, foi gradual (ver sua linha de tempo na Ilustração 8). As primeiras escolas envolvidas por essa ação aplicaram provas para alunos de 1ª série em 2001. Em 2002, esses mesmos alunos, quando aprovados, fizeram provas de 2ª série nas três primeiras unidades letivas; em 2003, provas de 3ª série e, em 2004, provas de 4ª série. Por essa razão, o período 2001 – 2004 foi considerado um ciclo completo da política.

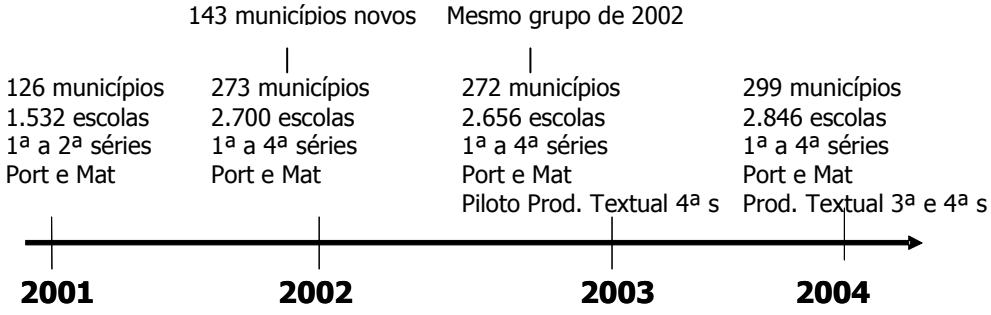


Ilustração 8: Linha de tempo da Avaliação da Aprendizagem (AA) quanto à entrada de municípios, escolas, séries e disciplinas avaliadas.

O fechamento do ciclo permitiu, pela primeira vez em 2004, o cruzamento dos dados da AD com a implementação de um ciclo de provas da AA porque, nesse ano, os alunos que participaram das

escolas da AA desde 2001, sem reprovação ou abandono, estariam fazendo a prova de AD na 4ª série, como exemplificado pelo "aluno João" na Ilustração 9 a seguir. O ciclo da AA pode, nesse caso, ser acompanhado com um indicador externo para o desempenho em Língua Portuguesa e Matemática, além dos dados do Censo Escolar do MEC.

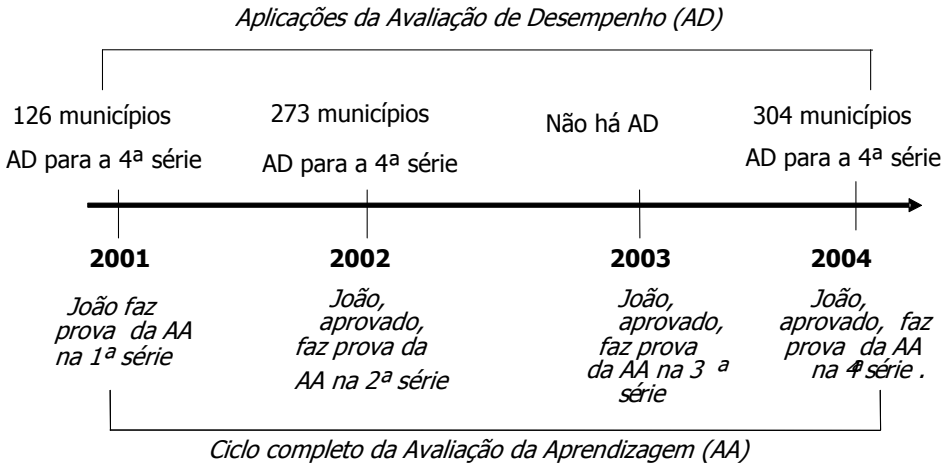


Ilustração 9: Representação, a partir do "aluno João", de um ciclo completo da *Avaliação de Aprendizagem*, cruzado com informações sobre a *Avaliação de Desempenho* (AD) paralela nesse mesmo período.

O detalhamento da política de AA, feito nessa subseção, aponta para a centralidade da escola na sua implementação, razão pela qual se tornou o foco, como *stakeholder* principal, da investigação em relato. A conclusão do primeiro (e único) ciclo da política, suas características (que a tornaram diferente das políticas implementadas à época no Brasil), e a existência de indicadores externos que pudessem ser utilizados para a compreensão dos usos feitos e para a percepção da utilidade foram a base para a escolha da AA como foco da pesquisa.

Em linhas gerais, essa subseção detalhou o foco da pesquisa em relato e o contexto no qual foi formulado e implementado. A próxima subseção, Lógica da Pesquisa, apresenta o quadro de análise e o quadro operacional utilizados para levantar as contribuições da Avaliação de Aprendizagem.

3.2 A lógica da pesquisa

Esse estudo investigou as contribuições de políticas de avaliação educacional, quando implementadas em larga escala, para as escolas públicas. Entendeu como contribuições o conjunto de elementos de Utilidade e Uso, discutidos a partir de base teórica apresentada na Seção 02. Para a investigação, escolheu como foco a política de Avaliação da Aprendizagem (AA), implementada pelo Governo do Estado da Bahia junto às redes estadual e municipais (parceiras do Estado) no período de 1999 a 2004 (com primeira aplicação de provas em 2001)[52], descrita na subseção anterior. A investigação partiu da seguinte pergunta: **quais as contribuições de um ciclo completo da política de Avaliação da Aprendizagem para as escolas por ela envolvidas?**

Derivada do quadro teórico, a hipótese trabalhada foi: em políticas de avaliação educacional em larga escala, os resultados são elementos pouco utilizados e é o acontecimento da avaliação que afeta as instituições em nível micro (escolas).

A pesquisa buscou inicialmente levantar, através dos itens verificadores da categoria Utilidade, a presença de características uso-conducentes no delineamento e implementação da política. De acordo com o modelo de meta-avaliação do JCSEE e do *checklist* de meta-avaliação elaborado por Stufflebeam (1999), quando presentes essas características apontariam para a qualidade da avaliação e para a concretização dos seus usos. Na categoria Utilidade adaptada para o presente estudo, foi incluída a percepção de utilidade como elemento atrelado à qualidade e ao uso. Por essa razão, foi feito o levantamento dos relatos sobre a percepção das escolas sobre a utilidade da AA.

---

52 O autor do presente estudo foi coordenador do Projeto de Avaliação Externa no período de 1999 a 2004, o que facilitou enormemente o acesso aos dados. O referido projeto, sob direção da UFBA (ISP), instaurou, desde agosto de 2001, um Comitê Científico que tinha, dentre suas atribuições, "aprovar solicitações oriundas de pessoas externas à Agência interessadas no uso dos dados por ela gerados" e "aprovar trabalhos produzidos por membros da equipe da Agência que se destinam à divulgação pública dos dados por ela gerados". As bases de dados oriundas das atividades do Projeto de Avaliação Externa foram autorizadas pelo Comitê e utilizadas para investigação de um trabalho de graduação, cinco dissertações de mestrado e três teses de doutorado até o momento. Dentre elas, está o presente estudo. Além das bases de dados e dos documentos oficiais do Projeto, este autor utilizou, para construção da sua tese, relatórios técnicos internos de acesso exclusivo da equipe central da avaliação, dos técnicos da SEC e da direção do ISP. Para tanto, assumiu o compromisso de sigilo sobre os dados individuais das escolas envolvidas pela avaliação externa. O autor agradece à equipe do Projeto de Avaliação Externa. Agradece especialmente a Carlos Henrique Nunes, psicometrista do Projeto de Avaliação Externa responsável pelas análises psicométricas da Avaliação da Aprendizagem e pela Avaliação de Desempenho em 2004; a Luis Fernando Pithon Sarno, coordenador de aprendizagem do Projeto e responsável pelo tratamento dos RD e elaboração dos relatórios síntese em 2004; a Olívia Maria Silveira, coordenadora de administração de instrumentos do Projeto; e a Rosana de Freitas Castro, professora pesquisadora da UFBA que utilizou as bases de dados da Avaliação Externa no seu doutoramento. Essas quatro pessoas coordenaram a construção e aprimoraram as bases de dados da Avaliação Externa utilizadas no presente estudo. As bases do Censo Escolar foram obtidas por meio de solicitação formal à Secretaria da Educação da Bahia, respondidas pela remessa de um CD com as bases de 2001 a 2005. O autor agradece a Mariano Romário Lima e Ilza Patrícia Carvalho, técnicos da SEC responsáveis pela montagem da base de dados solicitada.

Em um segundo momento, a investigação concentrou-se nos usos feitos. Inicialmente, focalizou o uso da avaliação para a tomada de decisões, sendo esse o uso instrumental clássico. Para tanto, buscou os relatos de usos feitos pelas escolas públicas, encaminhados à equipe central da avaliação por meio dos Relatórios do Diretor (RD). Nesse momento, expandiu a consulta sobre usos para além dos resultados, incluindo outros elementos da avaliação. Em seguida, voltou-se para o atingimento dos objetivos da política, na busca por uma relação entre o uso e a finalidade da política. Ao fazê-lo, lançou mão das variações nas taxas oficiais (aprovação, reprovação, abandono e proficiência em português e em matemática). Essas duas etapas maiores averiguaram a primeira parte da hipótese: os resultados são elementos pouco utilizados. Em seguida, o estudo buscou usos conceituais, na verificação da segunda parte da hipótese: é o acontecimento da avaliação que afeta as instituições por ela envolvidas. A figura a seguir ilustra a lógica da pesquisa:

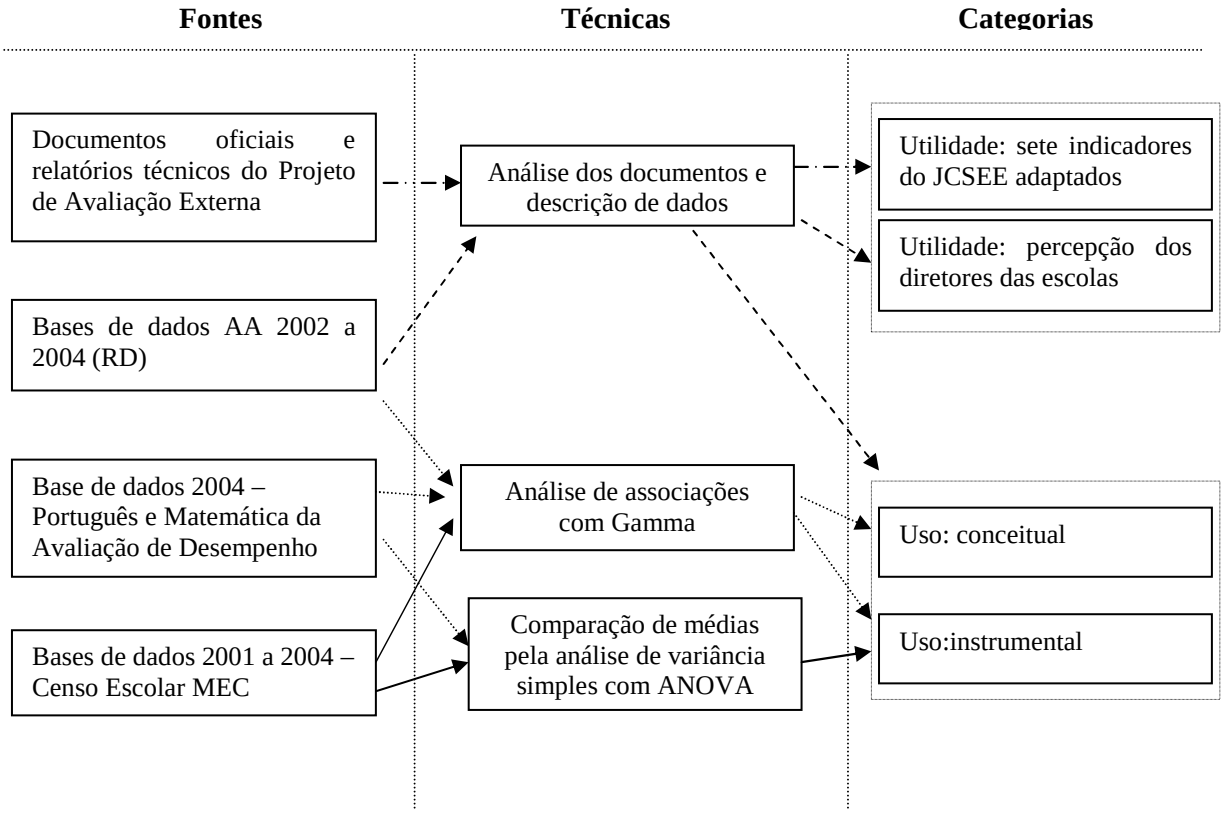


Ilustração 10: Representação do desenho da pesquisa sobre as contribuições da AA.

Os documentos oficiais do Projeto de Avaliação Externa foram as fontes de informações que permitiram o levantamento dos objetivos e das características da AA e dos elementos que respondem aos sete indicadores iniciais da categoria Utilidade. Através da análise desses documentos foi também possível delinear o contexto no qual essa política esteve inserida quando da sua formulação e implementação. Tal delineamento foi feito no sentido de permitir a análise posterior, em contexto de política pública, das contribuições (ou sua ausência) relatadas pelas

escolas. Os Relatórios Síntese[53] e as bases de dados construídas após tratamento dos RD foram as fontes para os relatos sobre percepção de utilidade e usos concretizados (instrumentais, conceituais e rituais) da AA. As bases de dados da Avaliação de Desempenho (AD) e aquelas do Censo Escolar permitiram a análise do atingimento dos objetivos da AA após o levantamento de dados sobre seu uso para a tomada de decisões (no contexto da escola pública, relativa ao planejamento do curso e das aulas).

Na busca dos usos concretizados, foram conduzidas comparações e associações entre os comportamentos identificados a partir de um dado de contexto: o ano de envolvimento do município, no qual se localizavam as escolas, com a política de avaliação. Como a expansão da avaliação foi gradual na Bahia, um grupo grande de municípios foi envolvido no início da implementação da política (e chegou até 299 no último ano) e outro grupo teve um envolvimento curto ou não teve relação com a AA. A diferença facilitou o contraste entre os grupos. Para alguns indicadores, foi feita uma associação dos comportamentos apresentados pelos grupos com os resultados da AD2004, com objetivo de identificar alguma tendência que pudesse ser relacionada a um desempenho mais alto em Português e/ou Matemática. Um exemplo: buscou-se saber se as escolas teriam utilizado os elementos da AA para o planejamento do curso em 2004. Em seguida, tentou-se estabelecer uma relação entre ter ou não planejado 2004 com dados da AA e o tempo de envolvimento com a política de avaliação. Por fim, investigou-se se os comportamentos observados (planejou x não planejou) guardavam alguma relação com o desempenho dos alunos de 4ª série em Português e Matemática.

As bases de dados foram construídas com variáveis em três níveis: nominais (tais como o nome do município), ordinais (como os tempos de envolvimento do município com a AA) e intervalares/escalares (a média de desempenho em Matemática, por exemplo). Por essa razão, optou-se pela análise de associação entre as categorias ordinais por meio de Gamma e da comparação das médias dos grupos com os resultados da AD pela análise da variância (ANOVA).

Esta subseção ofereceu o panorama geral da metodologia empregada no estudo em relato. A próxima subseção relaciona os documentos consultados e bases de dados utilizadas para que, em seguida, sejam detalhados os passos metodológicos.

---

53 Relatórios Síntese de RD, chamados mais tarde de Relatórios Síntese de Monitoramento, referem-se aos documentos técnicos produzidos pela equipe central da AA a partir da sistematização do monitoramento de cada uma das aplicações de prova. Em 2004, esses documentos passam a ser elaborados para distribuição com os *stakeholders* municipais. Para isso, o processo de tratamento de dados dos RD, após reformulação, passou a ser mais sintético e sistematizado e os RD foram alterados para a predominância de questões fechadas.

3.3 As fontes dos dados

Esse estudo utilizou, como fonte, documentos do Projeto de Avaliação Externa, agrupados em documentos gerais (aqueles da formulação e divulgação e os referentes à implementação), relatórios técnicos e relatórios do diretor (RD). Além disso, valeu-se das bases de dados existentes no Censo Escolar (MEC/SEC) e no Projeto de Avaliação Externa (UFBA/SEC). São eles:

3.3.1 Documentos gerais relativos à formulação da AA ou a sua divulgação:

- Projeto 06 do Educar para Vencer: Avaliação Externa (concepção do projeto no contexto do Educar para Vencer), 1999.
- Projeto de Avaliação Externa – SEC- UFBA/ISP – FAPEX – 1999 a 2004. Convênio 444/99 e seus aditivos.
- Relatório Final do Convênio 444/1999, elaborado pela UFBA/ISP como parte integrante da prestação de contas.
- *American Institutes for Research. Implementation of a statewide system for evaluation of student achievement. March*, 30, 2000.
- *American Institutes for Research. Implementation of a statewide system for evaluation of student achievement. June*, 15, 2000 (*final proposal*).
- Atas das reuniões do Conselho Consultivo do Projeto de Avaliação Externa.
- Regulamento do PIPEP – Programa de Incentivo aos pesquisadores do Ensino Público – Projeto de Avaliação Externa UFBA - PRPPG
- *Folder* promocional de o Programa Educar para Vencer, 2001.
- *Folder* Projeto de Avaliação Externa, 2003.
- *Folder* Avaliação da Aprendizagem: participação e contribuições das escolas 2002
- *Folder* Avaliação da Aprendizagem: participação e contribuições das escolas 2003
- Relatório Síntese de Monitoramento – AA 2004 1ª unidade.
- Relatório Síntese de Monitoramento – AA 2004 2ª unidade.
- Relatório Síntese de Monitoramento – AA 2004 3ª unidade.
- Relatório Avaliação de Desempenho 2004: resultados gerais e análises pedagógicas (SEC, 2005).
- PDE: orientações para implantação e implementação, 2001.
- Registros de reuniões entre os projetos da SEC, convites para participação em eventos promovidos por cada um e planos de eventos organizados em conjunto.
- Apresentações da AA em *powerpoint* para diversos públicos entre 2001 e 2004.
- Proposta da Educação Básica: Terra Bahia (SEC, 2002).
- Plano Estratégico SEC 1999-2002

3.3.2 Documentos referentes à implementação da política

- Manual do Diretor Avaliação da Aprendizagem 2002 – 1ª unidade
- Manual do Diretor Avaliação da Aprendizagem 2003 - 3ª unidade
- Manual do Diretor Avaliação da Aprendizagem 2004 – 3ª unidade
- Manual de Aplicação Avaliação da Aprendizagem 1ª série 2002 – 1ª unidade
- Manual de Aplicação Avaliação da Aprendizagem 3ª série 2002 – 3ª unidade
- Manual da Aplicação Avaliação da Aprendizagem 1ª série 2003 - 1ª unidade
- Manual da Aplicação Avaliação da Aprendizagem 4ª série 2003 – 3ª unidade

- Manual da Aplicação Avaliação da Aprendizagem 1ª série 2004 - 1ª unidade
- Manual da Aplicação Avaliação da Aprendizagem 4ª série 2004 – 3ª unidade
- Manual dos categorizadores de RD AA 2002 – 1ª ,2ª e 3ª unidades.
- Manual dos categorizadores de RD AA 2003 – 1ª ,2ª e 3ª unidades.
- Manual dos categorizadores de RD AA 2004 – 1ª ,2ª e 3ª unidades.
- Matriz de Referência Avaliação da Aprendizagem 1ª e 2ª séries
- Matriz de Referência Avaliação da Aprendizagem 3ª e 4ª séries
- Matriz de Referência Avaliação da Produção Textual 4ª série
- Manual de Revisão de Viés
- Documentos referentes aos processos licitatórios para impressão e distribuição dos materiais de AA 2001 - 2004
- Cartas-ofício encaminhadas às escolas antes de cada aplicação de provas (2002 – 3 unidades; 2003 – 3 unidades).
- Cartas-ofício encaminhadas às escolas em resposta às críticas e sugestões apresentadas por meio do RD (2002, 2003 e 2004).
- Correspondência eletrônica entre a coordenação da avaliação e representantes municipais.
- Correspondência eletrônica entre a coordenação da avaliação e representantes dos demais projetos da SEC.

3.3.3 Relatórios Técnicos (referentes a cada uma das aplicações dos testes da AA – 2002 a 2004)

- Relatório Análise psicométrica das provas de Português
- Relatório Análise psicométrica das provas de Matemática
- Relatório Determinação de linhas de corte para as escalas de proficiência de 1ª a 4ª séries da Avaliação de Aprendizagem
- Relatório Síntese da Logística
- Relatório Síntese da categorização e organização dos dados do Relatório do Diretor (até 2003)
- Relatório Síntese da categorização e organização dos relatórios do Coordenador de Aplicação

Não foram analisados os relatórios sobre os vídeos didáticos[54].

3.3.4 Relatórios do Diretor (formulários encaminhados às escolas em cada uma das aplicações)

- Relatório do Diretor (RD) 2001 – 2004

Ao todo, foram 11 formulários RD diferentes entre 2001 e 2004. Esses documentos dispuseram de um bloco de questões comuns, repetidas ao longo das aplicações, e de questões específicas a uma determinada unidade letiva. Dentre as questões comuns, estiveram aquelas relativas a problemas de distribuição dos materiais e atrasos nas entregas, às dificuldades observadas na aplicação dos testes, às questões de linguagem dos manuais e testes e às ações definidas pela comunidade escolar para o combate aos problemas porventura diagnosticados. No bloco de questões

54 Os relatórios sobre utilização dos vídeos didáticos não foram considerados no presente estudo. Apesar de serem componentes da AA, os vídeos eram encaminhados às escolas com abordagens didáticas para endereçamento dos principais problemas identificados em Português e Matemática, não fazendo parte dos elementos da avaliação propriamente dita.

específicas estiveram, por exemplo, as perguntas que buscaram a percepção das escolas sobre a utilidade da AA ou sobre a utilização dos seus resultados do ano anterior para o planejamento do ano em curso.

3.3.5 As bases de dados

- Base de categorização dos RD na AA 2002, nas três unidades, em Excel.
- Base de categorização dos RD na AA 2003, nas três unidades, em Excel.
- Base de categorização dos RD na AA 2004, nas três unidades, em SPSS.
- Base AD 2004 completíssima (com dados da AD 2002 e do questionário do diretor em 2004), em SPSS.
- Bases com as taxas de aprovação, reprovação e abandono (2001 a 2005), fornecidas pela SEC, em Excel.

Todas as bases AA foram resultantes do processo de categorização e tabulação dos RD. A base AD 2004 foi composta com os dados da aplicação da AD 2004: os resultados dos alunos de 4ª série em Língua Portuguesa e em Matemática e as respostas ao questionário do diretor, aplicado em paralelo às provas de 2004. Essa base precisou de um trabalho de nomeação a partir de dicionário disponibilizado. As demais foram recebidas já com os rótulos das variáveis.

A seguir são detalhados os passos metodológicos trilhados no presente estudo.

3.4 Os passos metodológicos

Com o objetivo de investigar as contribuições de um ciclo completo da AA, foram realizados os passos metodológicos[55] descritos a seguir.

3.4.1 Passo I: Sistematização do conjunto de documentos da AA e redução das bases de dados originais para as necessidades da investigação.

A primeira etapa do trabalho, ainda no início do doutorado, consistiu no levantamento dos documentos originais do Projeto de Avaliação Externa, especialmente aqueles que trataram da AA no contexto do Programa Educar para Vencer. Os documentos foram organizados cronologicamente e separados por natureza (documentos referentes ao delineamento da política, documentos de implementação da política e relatórios técnicos). Foram também buscados os modelos de RD e dos Manuais de Aplicação (professor e diretor) da AA de 2001 a 2004, de modo que se pudesse identificar, nas bases de dados categorizadas e tabuladas pela equipe central da Avaliação e nos relatórios síntese correspondentes, as questões que trataram de percepção de utilidade ou dos usos feitos com os resultados e o processo de avaliação. A partir da leitura desses documentos, foi possível o levantamento dos objetivos e características da AA. Para a descrição do contexto no qual houve a implementação da política, foi utilizado o material já relacionado para a dissertação de mestrado desse autor.

Ainda nessa fase, uma segunda etapa de tratamento das bases de dados disponíveis foi realizada após a qualificação. Para tanto, utilizou-se o SPSS. Após redução das imensas bases da AA2004 e da AD 2004 (resultados de alunos e respostas dos diretores, com mais de 1.000 variáveis no total) aos indicadores que tratassem de uso ou de utilidade, foram nomeadas as variáveis da base AD2004 a partir de dicionário fornecido pela equipe central. As bases da AA2004 já tinham seus respectivos rótulos.

Além disso, nessa etapa foram investigados os manuais de categorização dos RD, utilizados na capacitação dos consultores responsáveis pela tabulação dos dados, para análise das informações nas bases de dados e da qualidade dos relatórios síntese de monitoramento. Após análise dos dados da AA de 2002 e 2003, optou-se por utilizar as informações já sintetizadas pela equipe da avaliação e disponíveis nos Relatórios Síntese das aplicações por unidade letiva (1ª, 2ª e 3ª, com

ênfase na 3ª unidade) e descartar o trabalho direto com essas duas bases. Por fim, por falta de padronização no trabalho de tabulação dos dados, optou-se pela não utilização das bases com dados oriundos dos RD da AA 2001.

É importante ressaltar que o *feedback* das escolas derivado dessas bases ou obtido dos Relatórios Síntese não representa o conjunto de unidades escolares abrangidas pela AA. Há na amostra um viés: as escolas que encaminharam os RD dentro do prazo, de alguma maneira, distinguiram-se das demais por cumprir os prazos da AA e manter a comunicação escola – equipe central (ou por terem estado ligadas a coordenações municipais que o fizeram). Houve, no entanto, opção pela sua utilização no presente estudo porque os registros dessas escolas permitem esclarecer o que ocorreu durante a implementação da AA e o uso de seus materiais por aquelas escolas que cumpriram o calendário proposto. Em um delineamento ideal de pesquisa, as escolas (e mesmo as redes municipais) não respondentes ao longo do ciclo da AA deveriam ser acompanhadas. Entretanto, isso não foi possível no contexto da AA ou do presente estudo, por razões de logística, custo e tempo. A tabela a seguir oferece o panorama de RD respondidos e encaminhados à equipe central, ao longo dos anos, que, após categorização e tabulação, formaram as bases de dados da AA.

Tabela 1: Panorama de RD enviados às escolas e encaminhados, depois de respondidos, de volta à equipe central da Avaliação entre 2001 e 2004.

Relatórios do Diretor	**Ano**			
	2001	**2002**	**2003**	**2004***
Total geral de RD enviados às escolas / ano	3.064	8.100	7.968	8.264
Total de RD respondidos pelas escolas / ano	1.500	4.600	4.590	3.948
Percentual de RD respondidos (aproximado)	49%	57%	58%	48%

Fonte: Relatório de Conclusão do Convênio 444/99 e Relatório Síntese de Monitoramento AA 2004-3ª unidade

No grupo de escolas respondentes, há unidades que enviaram os três relatórios por ano ou que enviaram os RD em uma ou outra unidade. De maneira geral, comparados os envios em um mesmo ano, percebe-se uma queda na freqüência de encaminhamento do RD na 3ª unidade letiva (DANTAS, 2005). Ao longo dos quatro anos do primeiro ciclo da AA, a tabela acima mostra que o ano de 2003 foi aquele com maior percentual de respostas pelas escolas e que esse percentual caiu em 10% quando comparado a 2004. Essa queda pode ter sido devida à entrada de novos municípios em 2004 ou ainda a um "cansaço" no envio dos RD pelas escolas que o fizeram em

55 A definição dos passos metodológicos foi feita a partir da leitura de LAVILLE, Christian e DIONNE,Jean (1999) e QUIVY, Raymond e CAMPENHOUDT, Luc van (1998).

anos anteriores. À exceção dos esclarecimentos diretos sobre dúvidas quanto ao processo de AA, não havia um benefício direto para os respondentes.

Finda a organização das bases AA e AD, foram trabalhadas – isoladamente - as bases oriundas do Censo Escolar MEC/SEC para a verificação do atingimento dos objetivos da AA. Nesse caso, por razões operacionais, optou-se pelo trabalho com as bases de 2001 e 2004 (entrada e finalização do ciclo da AA) e, em mais uma etapa de redução, a análise foi concentrada em duas séries do Ensino Fundamental, 1ª e 4ª. A 1ª série foi escolhida por ser a porta de entrada no Ensino Fundamental e aquela considerada crítica por muitos. Já a 4ª série foi escolhida por ser entendida como a finalização da primeira etapa do Ensino Fundamental (1ª a 4ª série) e, no caso em tela, do primeiro ciclo da AA. Novamente, após redução das bases para os indicadores do Quadro Operacional, foi feito uma fusão entre 2001 e 2004 e, em seguida, foram criadas novas variáveis que trataram da diferença entre as taxas de 2004 e aquelas apresentadas pela mesma escola em 2001. As bases do Censo Escolar são construídas a partir das respostas das escolas e não há uma verificação sobre a veracidade desses dados. Por essa razão, as análises oriundas dessas bases fornecem apenas um panorama geral das tendências das redes.

Nas bases finais de trabalho, as escolas foram agrupadas pelo período de envolvimento do município onde se localizam com o Programa Educar para Vencer e, mais especificamente, com a AA. Os primeiros municípios estabeleceram parcerias em 1999/2001. Um novo grupo foi adicionado ao já existente em 2002, que se manteve estável em 2003. Esse grupo foi identificado como 2002/2003. Em 2004, novos municípios estabeleceram parceria no início do ano (nominados AA 2004) e um grupo menor foi envolvido apenas para a AD2004. Dessa categorização surgiram dois grandes grupos: aqueles que deveriam ter sido impactados pela AA (os que entraram em 2003 ou anos anteriores) e aqueles que, pressupunha-se, não teriam tido tempo para sofrerem um impacto (aqueles envolvidos em 2004 ou não envolvidos). Os dados sobre a expansão da AA são detalhados na primeira subseção da Metodologia.

3.4.2 Passo II: Criação do Quadro de Pesquisa: Quadro de Análise e Quadro Operacional

Em paralelo ao Passo I e até o final da pesquisa, buscou-se na literatura o fundamento para a construção do Quadro de Análise e, a partir dos documentos e bases existentes sobre e oriundos da AA e AD, do Quadro Operacional. O marco teórico foi fundamentado na seguinte lógica: contexto mais amplo: política pública; conceito mais abrangente: avaliação; primeiro recorte: avaliação educacional; segundo recorte: qualidade da avaliação investigada pela meta-avaliação

(ver mapa conceitual no Apêndice 1). A pesquisa, a partir daí, focalizou as contribuições das políticas de avaliação, adaptando, inicialmente, a categoria Utilidade do JCSEE, detalhada pelos itens verificadores propostos por Stufflebeam (1999) para avaliação de programas educacionais. Nessa etapa, buscou discutir a aplicabilidade dos itens verificadores à meta-avaliação de políticas de avaliação educacional, com uma proposta de enriquecimento da categoria pela introdução do 8º indicador, percepção da utilidade. Em seguida, a partir do panorama de estudos sobre usos da avaliação, desenhado na aparente oposição de Carol Weiss a Michael Patton e na classificação de usos em instrumental e conceitual, propôs-se a segunda categoria de análise, Uso. Essa segunda categoria foi dividida em Uso Instrumental e Uso Conceitual.

Como a pesquisa utilizou dados secundários e documentos do Projeto de Avaliação Externa, o Quadro de Análise foi operacionalizado como mostram os dois quadros a seguir (Quadro 7 – Categoria Utilidade e Quadro 8 – Categoria Uso).

Quadro de Análise*		**Quadro Operacional**		
Indicadores	**Dimensões**	**Fonte de dados**	**Abordagem metodológica**	**Tempo de coleta considerado**
U1	Identificação dos *stakeholders*	Documentos da AA; relatórios técnicos de avaliação.	Análise documental	2000 – 2003
	Identificação do atendimento da demanda	Documentos da AA; RS; *folder*s de conclusão AA/ano.	Análise documental; análise de tendências nas bases.	2002-2003
U2	Credibilidade percebida	Declaração do cliente; RS.	Análise documental	1999 – 2003
U3	Escopo	Documentos da AA; matrizes de referência; RS.	Análise documental	2000 – 2004
	Coleta	Relatórios técnicos da AA.	Análise documental	2001 – 2004
U4	Bases para a definição do valor	Relatórios técnicos da AA (Angoff); RS.	Análise documental	2001 – 2004
	Valores definidos	Relatórios técnicos da AA.	Análise documental	2001 – 2004
U5	Forma dos relatos	Documentos de comunicação utilizados pela AA.	Análise documental	2001 – 2004
	Conteúdo dos relatos	Documentos de comunicação utilizados pela AA.	Análise documental	2001 – 2004
U6	Perecibilidade	Relatórios técnicos da AA; RS.	Análise documental	2001 – 2004
	Disseminação	Relatórios técnicos da AA; RS.	Análise documental	2001 – 2004
U7	Impacto	Documentos de comunicação utilizados pela AA.	Análise documental	2003 – 2004
U8	Percepção de Utilidade	Base de dados da AA2004 – 3ª unidade; RS 2003 3ª unidade.	Análise de tendências (respostas ao RD); análise documental.	2004

Legenda: RD – Relatório do Diretor; RS – Relatório Síntese; AA – Avaliação de Aprendizagem
* Os itens de verificação estão detalhados na subseção 2.4.2.

Quadro 7: Quadro Operacional para a categoria Utilidade

Quadro de Análise			Quadro Operacional		
Indicadores	**Dimensões**	**Itens de verificação**	**Fonte de dados**	**Abordagem metodológica**	**Tempo de coleta considerado**
Instrumental	Tomada de decisões	Utilização dos resultados da AA para planejamento.	Documentos da AA, base de dados AA 2004 (1ª e 2ª unidades) e base de dados AD 2004.	Análise dos documentos; análise do posicionamento das escolas sobre o uso da AA para a tomada de decisões/planejamento; cruzamento dos dados de planejamento com os resultados (Gamma e ANOVA).	2004
		Utilização de outros elementos da AA para planejamento.			
	Atingimento dos objetivos da AA	Dif. tx aprovação 1ª e 4ª séries do EF	Base de dados do Censo Escolar 2001 - 2004 (1ª e 4ª séries).	Busca da variação das taxas e cruzamento (Gamma e ANOVA) com dados de envolvimento na AA.	2001 e 2004
		Dif. tx reprovação 1ª e 4ª séries do EF			
		Dif. tx abandono 1ª e 4ª séries do EF			
		Média *Theta* Matemática 4ª série	Base de dados da AD 2004	Comparação de médias entre grupo de envolvimento com a AA e desempenho (ANOVA).	2004
		Média *Theta* Português 4ª série			
Conceitual	Político-persuasório	Identificação de necessidade de capacitação.	Documentos da AA, base de dados AA 2004 (3ª unidade).	Análise dos documentos; análise do posicionamento das escolas sobre o uso da AA para aspectos não voltados ao planejamento.	2004
		Envolvimento dos pais.			
		Monitoramento de professores (pressão).			
	Motivacional	Efeito motivacional para os alunos			
	Partilha	Discussão e entendimento coletivos das questões do ensino.			

Quadro 8: Quadro Operacional para a categoria Uso

A construção do quadro de análise e do quadro operacional foi refinada no desenrolar da pesquisa, tanto pelas contribuições da teoria quanto pelas limitações dos dados existentes, e somente assumiu as feições colocadas nos dois quadros acima ao final do trabalho. Com os dados organizados e a base teórica estabelecida, foram conduzidos os demais passos do estudo.

3.4.3 Passo III: Análise dos documentos da AA para resposta aos itens verificadores dos sete primeiros indicadores da categoria Utilidade (U1 a U7).

O presente trabalho aproximou os padrões da categoria Utilidade do JCSEE e itens de verificação do *checklist* de Stufflebeam (1999) ao estudo das contribuições (usos e utilidade) de políticas de avaliação e os aplicou a uma delas, a Avaliação de Aprendizagem, implementada pelo Governo do Estado da Bahia no período de 2001-2004. A proposta original do referido *checklist* incluiu uma pontuação por categoria e a definição dos itens essenciais na determinação de uma avaliação de qualidade, ainda que não tivesse apresentado uma linha de corte que distinguisse entre o aceitável e o não aceitável, por categoria ou de maneira global.

O estudo em relato não se valeu dessa pontuação, visto que vários dos itens verificadores não poderiam ser aplicados diretamente sobre políticas públicas de avaliação. Além disso, o objetivo da pesquisa não foi determinar se a AA era ou não uma experiência de qualidade e sim o levantamento das suas contribuições. A análise conduzida sobre a AA, nos sete primeiros indicadores da categoria Utilidade, resultou na confirmação da presença ou da ausência dos elementos apontados pelos itens verificadores aplicáveis. Para tanto, foram consultados os documentos da AA de 2001 a 2004, incluindo apresentações sobre o Projeto de Avaliação Externa em *power point*, ofícios e *mails* trocados entre a coordenação do referido Projeto e vários *stakeholders*, manuais e relatórios técnicos. Cada indicador foi observado a partir dos itens verificadores adaptados e, para apresentação dos resultados, foi construído um quadro síntese por indicador, no qual eram marcados os itens atendidos pela AA.

Para síntese final do comportamento da AA nos sete indicadores, propôs-se então uma escala de probabilidade de uso por indicador, em três níveis: probabilidade alta, quando mais de 70% dos itens verificadores no indicador estiveram presentes; média, quando estiveram presentes entre 41 e 70% dos itens verificadores; e probabilidade baixa, quando 40% ou menos dos itens verificadores foram observados. O quadro a seguir apresenta essa escala.

Presença dos itens por indicador	Escala de probabilidade de uso
Até 40%	Baixa
41 a 70%	Média
Mais de 70%	Alta

Quadro 9: Níveis da escala de probabilidade de uso

3.4.4 Passo IV: Análise dos documentos e bases da AA para resposta à dimensão Percepção de Utilidade (U8) da categoria Utilidade.

Para análise da percepção de utilidade relatada pelas escolas, foi feita uma busca pelos formulários de RD nos quais, em algum momento da implementação da AA, houvesse perguntas direcionadas a coleta desse dado. De maneira sistematizada, em duas ocasiões foram feitas perguntas diretas: na 3ª unidade de 2003 e, novamente, na 3ª unidade de 2004.

Na terceira unidade de 2003, o RD incluiu uma grade com 20 áreas (apresentada na seção Resultados, Subseção 4.1.9) para as quais a escola deveria colocar sua percepção do efeito da AA em cada uma das quatro séries do Ensino Fundamental menor. Dentre essas áreas estavam: qualidade geral do ensino, planejamento do curso, comunicação escola – pais, comunicação coordenação – professores, por exemplo. Optou-se pela utilização das informações divulgadas no Relatório Síntese correspondente, já que o mesmo trazia esses dados tratados, categorizados e tabulados. Naquela unidade foram envolvidas 2.623 escolas urbanas (905 estaduais e 1.718 municipais) de 272 municípios. Desse conjunto, 1.465 unidades (427 estaduais e 1.038 municipais, aproximadamente 56% do total inicial) encaminharam os RD a tempo de processamento para a composição do Relatório Síntese, mas apenas 290 manifestaram sua opinião sobre o efeito da AA por série. Os dados dessas 290 escolas, apesar da não representatividade, foram considerados neste estudo pela variabilidade da percepção quando observadas as diferentes respostas nas quatro séries do Ensino Fundamental.

Além da grade, os diretores foram convidados a responder uma questão aberta (Q. 9), transcrita abaixo, justificando sua escolha sobre as três áreas nas quais, para eles, o efeito da AA (negativo ou positivo) havia sido maior.

> Q 9. Use o espaço abaixo para justificar sua resposta sobre três áreas (do quadro anterior) onde sua escola julga que a Avaliação de Aprendizagem tem produzido os maiores efeitos (positivos ou negativos). Escreva outras observações que considerar pertinentes.
> Caso o espaço não seja suficiente, anexe uma página ao relatório e continue.
>
> RD 2003 – 3ª unidade

Novamente, foram utilizadas as informações contidas nos Anexos do Relatório Síntese 2003 – 3ª unidade para ilustrar as respostas encontradas na grade.

Para as informações oriundas da AA 2004-3ª unidade, optou-se pelo trabalho direto na base de dados correspondente. Na ocasião, o RD incluiu um bloco de questões fechadas que buscavam levantar a percepção de efeito da AA na escola junto aos professores das quatro séries iniciais (1ª

a 4ª do EF). Para a percepção de utilidade, foram escolhidas as seguintes perguntas (repetidas nas 4 séries):

> Q.18 Perguntas para os professores da 1ª série:
>
> - A equipe escolar percebeu uma melhoria no aprendizado de Língua Portuguesa e Matemática em seus alunos de 1ª série, como efeito do trabalho com a Avaliação de Aprendizagem?
> Língua Portuguesa: Sim ( ) Não ( ) Não é possível afirmar ( )
> Matemática: Sim ( ) Não ( ) Não é possível afirmar ( )
>
> - Em 2004, a equipe escolar percebeu que os alunos da 1ª série tornaram-se mais familiarizados com o formato dos testes da Avaliação da Aprendizagem?
> Sim ( ) Não ( ) Não é possível afirmar ( )
>
> RD 2004 – 3ª unidade

Na 3ª e 4ª série, a primeira pergunta incluiu também Produção Textual. Além da questão sobre a AA e a aprendizagem, os professores foram perguntados se percebiam uma maior familiarização com o formato da prova, já que haviam manifestado, em RDs anteriores, que o alunado não demonstrava um bom desempenho porque não conhecia o formato usado pela AA.

Na 3ª unidade de 2004, a AA aplicou testes em 299 municípios, nos quais foram envolvidas 2.846 escolas urbanas (2.087 da rede municipal e 759 da rede estadual). Dessas, 1.200 unidades (269 estaduais e 931 municipais, 42% do total) preencheram e encaminharam o RD a tempo do processamento que resultou no Relatório Síntese e na base de dados utilizada para a presente pesquisa. Dos 1.200 RD processados, a grande maioria respondeu às perguntas sobre percepção. Após a fusão da base de dados AA com AD 2004 e da exclusão dos municípios de Salvador, Nilo Peçanha e Jacobina da base de dados, o número válido final de escolas foi reduzido para 917. Esses três municípios foram retirados da base porque, diferentes em relação aos demais, apenas participaram da AA com as escolas estaduais. Sendo discrepantes, optou-se pela sua exclusão da base de dados para o presente relato.

Foi feita uma análise da freqüência das respostas por série, tanto em relação ao efeito da AA sobre a aprendizagem como da acomodação dos alunos com o formato dos testes. Por fim, do Relatório Síntese de Monitoramento da 3ª unidade de 2004 foi retirada uma tabela com os dados de proficiência dos alunos naquela unidade, na busca por elementos que ajudassem a entender o comportamento das escolas nas respostas dadas sobre o efeito da AA nas quatro séries.

Os passos III e IV permitiram o levantamento dos indicadores na categoria Utilidade. A partir desse momento, o foco saiu da categoria Utilidade e voltou-se para a categoria Uso, composta por

dois indicadores: Uso Instrumental e Uso Conceitual. Os passos seguintes foram dedicados à análise de suas dimensões.

3.4.5 Passo V: Análise do uso da AA para a tomada de decisões (Uso Instrumental)

De acordo com o Quadro de Análise, o Uso Instrumental foi verificado por meio de duas dimensões: uso para tomada de decisões e atingimento dos objetivos da política. O Passo V trata da primeira.

Para análise da utilização da AA para a tomada de decisões, optou-se por concentrar as análises na base de dados da AA 2004 – 1ª, pela presença de uma pergunta direta sobre o tema, e descartar as demais. A base individual já havia sido reduzida na primeira etapa de pesquisa. Na etapa, foi feito uma fusão com a base AD 2004, para que houvesse identificação do ano de envolvimento do município e pudesse ser feita uma comparação de médias de desempenho dos alunos da 4ª série com o comportamento adotado pela escola em relação ao uso da AA. O Relatório Síntese da AA 2004 – 1ª unidade informa que 2.567 escolas em 272 municípios foram envolvidas na AA (783 municipais e 785 estaduais). Naquela unidade, 1.349 escolas (aproximadamente 53% do total) encaminharam o RD de volta para a equipe central a tempo de serem processados. Após a retirada dos municípios de Salvador, Nilo Peçanha e Jacobina e do cruzamento com a base AD, a base final resultante contava com 1.013 escolas (39% do total).

Como se tratava da AA na primeira unidade, o RD correspondente continha uma questão, a de nº 16, transcrita a seguir, sobre a utilização dos resultados de 2003 para o planejamento de 2004. As alternativas, além dos resultados, lidavam com as matrizes de referência, além de deixar um campo em aberto, para que a escola relacionasse outros elementos da AA que tivessem utilizado para planejar. Como o diretor da escola poderia marcar mais de uma alternativa na resposta, a equipe central processou cada alternativa como uma questão isolada, com as respostas Sim (marcou) ou Não (não marcou).

16. Marque com um × os materiais que sua escola utilizou para o planejamento de curso em 2004.
No planejamento para o ano de 2004, sua escola:
( ) Utilizou os diagnósticos das turmas, obtidos nas três unidades de 2003.
( ) Utilizou os diagnósticos das turmas, obtidos **apenas** na 3ª unidade de 2003.
( ) Utilizou as matrizes de referência de 1ª e 2ª séries.
( ) Utilizou as matrizes de referência de 3ª e 4ª séries.
( ) Utilizou a matriz de produção textual para a 4ª série.
( ) Utilizou outros materiais enviados pelo Projeto de Avaliação Externa.
Se sim, quais _____

RD 2004 – 1ª unidade

Foi feita uma descrição da freqüência das respostas obtidas nas seis alternativas da questão e um relato sobre as respostas no campo aberto. Posteriormente, fez-se uma comparação das médias (ANOVA) de *theta*-TRI em Língua Portuguesa e Matemática[56] com o comportamento das escolas no planejamento (sim e não para cada alternativa), de alguma maneira buscando identificar se essa utilização implicou, em tendência, um desempenho mais alto por parte dos alunos.

Concluída a busca por informações sobre a utilização da AA para a tomada de decisões, o estudo voltou-se para a análise do atingimento dos objetivos da política, ambos associados ao Uso Instrumental, detalhada nas duas subseções a seguir.

3.4.6 Passo VI: Análise do uso da AA para o atingimento dos objetivos da AA – Uso Instrumental

O segundo dos itens de verificação do indicador Uso Instrumental, Atingimento dos objetivos da AA, foi trabalhado em duas etapas: nesta, com a base de dados do Censo MEC/SEC, e a seguinte, com a base de dados AD2004. A base do Censo MEC/SEC permitiu o acompanhamento das variações nas taxas de aprovação, reprovação e abandono da 1ª e da 4ª séries (ano de entrada e ano final do Fundamental Menor, clientela afetada pela AA), como pode ser visto no quadro síntese a seguir.

Elementos	**Dados do Censo Escolar**		**Diferença entre as taxas 2004 - 2001**	**Variação das taxas**
Aprovação	Tx escolar 2001 (1ª e 4ª séries)	Tx escolar 2004 (1ª e 4ª séries)	Tx 2004 – Tx 2001	0 – Aprovação 2004 foi menor que a de 2001(série / escola). 1 – Não houve variação na taxa da série/escola. 2 – Tx 2004 foi maior que a de 2001(série / escola).
Reprovação	Tx escolar 2001 (1ª e 4ª séries)	Tx escolar 2004 (1ª e 4ª séries)	Tx 2004 – Tx 2001	0 – Reprovação foi maior em 2004(série / escola). 1 – Não houve variação na taxa da série/escola. 2 – Tx 2004 foi menor que a de 2001(série / escola).
Abandono	Tx escolar 2001 (1ª e 4ª séries)	Tx escolar 2004 (1ª e 4ª séries)	Tx 2004 – Tx 2001	0 – Abandono foi maior em 2004 (série/escola). 1 – Não houve variação na taxa da série/escola. 2 – Abandono em 2004 foi menor que em 2001 (série / escola).

Quadro 10: Indicadores utilizados para analisar o atingimento do objetivo da AA no presente estudo.

56 O *theta* é uma escala que funciona com uma média de zero e um desvio padrão igual a 1. Essa escala normalmente gera valores entre -4 e 4, com raros valores extremos. Entretanto, dado que o desempenho dos alunos de 4ª série, especialmente em Matemática, foi bastante insatisfatório, para a AD 2004 há valores extremos. Para interpretação do *theta*, os valores maiores indicam maior proficiência e os menores apontam para a baixa proficiência.

A partir das taxas de aprovação, reprovação e abandono, foram criadas duas novas variáveis: a diferença entre as taxas de 2004 e 2001 (tx 2004 – tx 2001) e sua variação (diminuiu, manteve, aumentou). As taxas de aprovação, por um lado, e de reprovação e abandono, por outro, têm sinais diferentes. Esperava-se, com o atingimento dos objetivos da AA, que a taxa de aprovação em 2004 fosse superior à taxa de 2001 (resultado positivo na equação tx 2004 – tx 2001). Da mesma forma, em 2004, esperava-se que as taxas de reprovação e o abandono tivessem caído quando comparados às taxas de 2001. Por isso, idealmente as tx 2004 – as tx 2001 teriam um resultado negativo.

Como já defendido no início do capítulo, os dados do censo utilizados referem-se aos anos de 2001 e 2004 exatamente por corresponderem ao início e ao final do primeiro ciclo da AA. O delineamento original da AA previa que apenas as escolas urbanas, nas redes municipal e estadual, estivessem envolvidas na avaliação bimestral. Para contraste, a base de dados 2001-2004 da presente pesquisa foi composta também com informações dos municípios que não fizeram parte da AA e das escolas rurais desses e daqueles que foram envolvidos na política. No total, foram sistematizados dados de 21.759 escolas de 414 municípios baianos (excluídos Salvador, Nilo Peçanha e Jacobina). Nessa composição, associou-se o código da escola ao município no qual se localizava e relacionou-se o município ao período de envolvimento com a AA, aferido pelo número de anos da parceria estado x município[57]. O principal interesse foi verificar se havia alguma tendência diferente quando se cruzavam dados da variação das taxas das escolas com aqueles de envolvimento do município com a AA. A base de dados oriunda dos bancos do Censo SEC/MEC foi, por essa razão, dividida pela localização das escolas, de modo que toda descrição e cruzamentos posteriores foram feitos a partir do *status* rural x urbano das mesmas.

Buscou-se assim contrastar o comportamento das escolas urbanas situadas em municípios envolvidos pela AA com escolas urbanas de municípios não envolvidos, por um lado, e contrastar escolas urbanas com escolas rurais dos mesmos municípios. A expectativa era de que quanto mais tempo a escola estivesse exposta à AA, tanto melhor estariam suas taxas. Como esclarecido anteriormente, foram definidas quatro faixas de envolvimento: 2001 e antes, 2002/2003, 2004 AA e 2004 AD. Para a análise de associações de dados ordinais, com poucas categorias, como foi o caso dos grupos por ano de envolvimento com a política de AA e das variações nas taxas entre

57 É possível que algumas escolas urbanas nesses municípios não tenham participado da AA ou que algumas escolas rurais o tenham feito. O Censo é construído a partir das informações encaminhadas pelas próprias escolas. Houve

2004 e 2001, utilizou-se o Gamma (γ) como medida (BABBIE, 1999; COSTA, 2004; LEVIN, 1987), tomando como nível de significância $\alpha \leq 0{,}05$. A lógica utilizada é ilustrada a seguir.

Uso Instrumental – Atingimento dos objetivos da AA

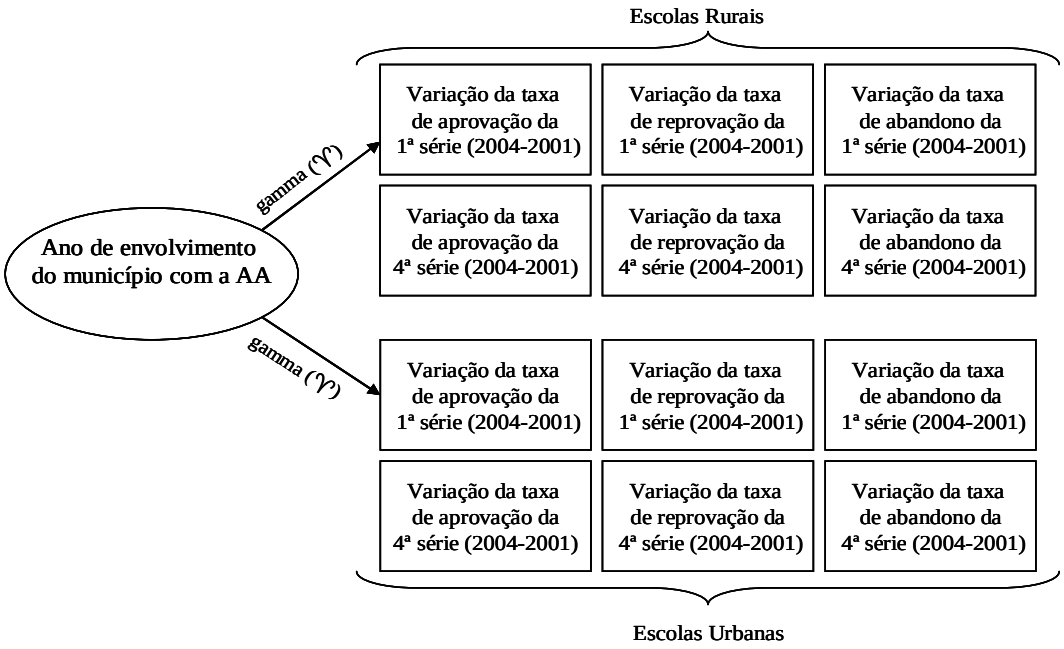


Ilustração 11: Lógica da associação entre o tempo de envolvimento dos municípios com a AA e a diferença das taxas de eficiência entre 2004 e 2001

Para análise de Gamma, observa-se que, quanto mais próximo de ±01 o resultado obtido, mais forte é a associação entre o tempo de envolvimento dos municípios com a política de AA e suas taxas de eficiência. Neste estudo, foram consideradas as associações medianas e fortes aquelas com Gamma > 0,50. As associações encontradas entre as taxas de aprovação, reprovação e abandono foram fraquíssimas ou não significativas. Por essa razão, optou-se por um segundo agrupamento das escolas em termos do envolvimento de seus municípios com a AA. Foram criados dois grandes grupos de escolas: aquelas localizadas nos municípios que não participaram da AA ou só iniciaram em 2004 (na AA ou mesmo apenas com a Avaliação de Desempenho, AD) e aquelas em municípios expostos ao ciclo completo ou parcialmente completo da AA. No primeiro grande grupo encontraram-se 33,3% das escolas e os 67% restantes estiveram associadas aos municípios do segundo grupo.

---

alguma variação do registro de localização da escola na base do Censo 2004 quando comparada à base do Censo 2001.

Foram contrastadas as médias escolares das diversas taxas e feita uma análise da variância (ANOVA), teste estatístico que visa verificar se existe uma diferença significativa entre as médias. A lógica dessa etapa está ilustrada a seguir.

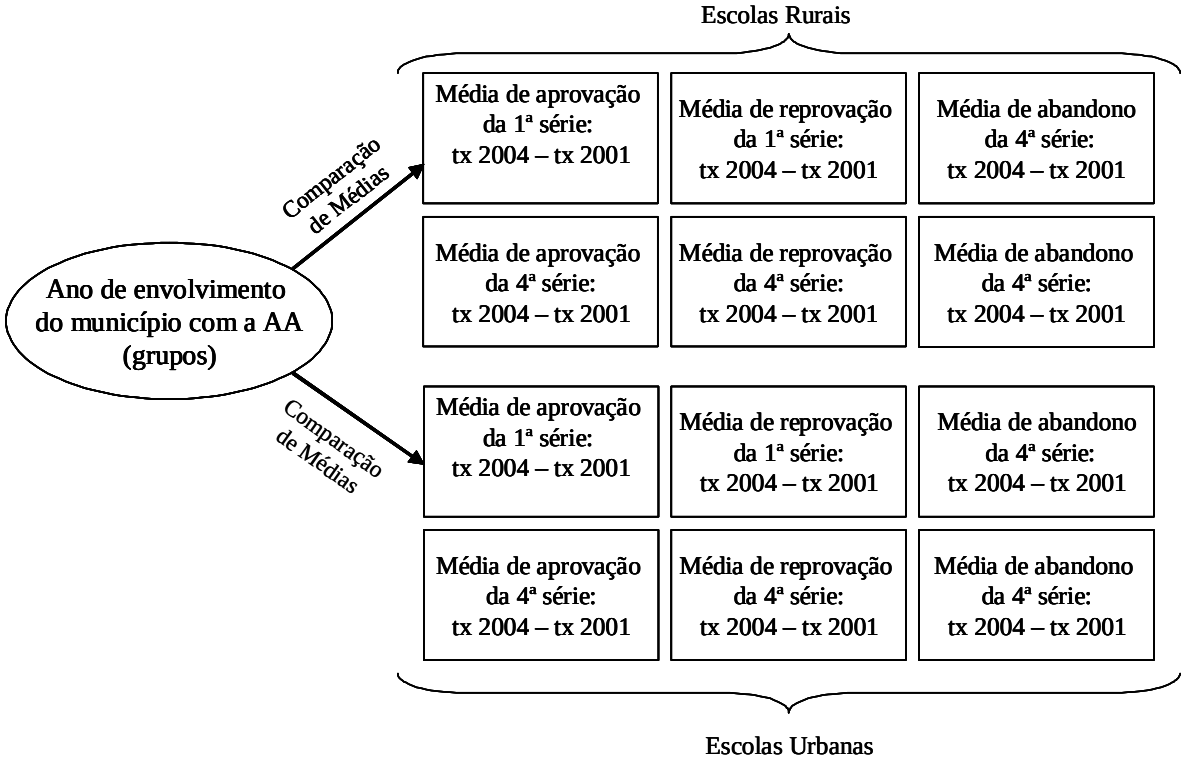


Ilustração 12: Lógica da associação entre os grupos por tempo de envolvimento dos municípios com a AA e a diferença das taxas de eficiência entre 2004 e 2001.

Concluída a etapa de trabalho com a base de dados oriunda do Censo, o estudo concentrou-se na base de dados da AD 2004, para verificação das variações relativas à proficiência do alunado em Língua Portuguesa e em Matemática.

3.4.7 Passo VII: Análise da base síntese da AD 2004 para levantamento das respostas sobre Uso Instrumental – indicador Atingimento dos objetivos da AA

A segunda etapa de análise do atingimento dos objetivos da política foi desenvolvida a partir da base de dados da AD 2004. No dia 25 de novembro de 2004, 2.808 escolas públicas da zona urbana de 304 municípios baianos associados ao Programa Educar para Vencer participaram da AD2004. Foram envolvidos 227.836 alunos de 4ª e 8ª séries do ensino fundamental e de 3ª série do ensino médio. Para a 4ª série, foram encaminhados materiais para 2.291 escolas (127.662 alunos). Houve, no entanto, uma evasão significativa (19,7% de escolas e 32,2% de alunos) quando consideradas as escolas que realmente participaram da AD. Dados brutos foram obtidos de 1.840 escolas e 86.548 alunos quando observada a 4ª série. De acordo com informações

divulgadas no relatório da AD 2004[58], contribuíram para estas perdas uma greve estadual e a aplicação tardia das provas, com impacto para escolas municipais que já tinham encerrado seus calendários de aulas.

Os resultados foram calculados, em sua etapa final, pelo valor de *theta* da Teoria da Resposta ao Item (TRI). Fez-se uma fusão da base AD 2004 com a preparada na etapa anterior, para que fosse possível agrupar as escolas por ano de envolvimento de seus municípios com a AA. Não foi necessário separar as escolas por localização (urbana x rural), visto que a AD só ocorria em escolas que tivessem registro como urbanas na base do Censo. Pelas razões já apontadas na subseção anterior, foram excluídos os dados dos municípios de Salvador, Jacobina e Nilo Peçanha. Após a análise das médias das escolas pelos períodos de envolvimento de seus municípios através do *Boxplot* no SPSS, da mesma forma que na fase anterior, optou-se pela análise da variância (ANOVA) entre as médias obtidas pelos dois grandes grupos de envolvimento de municípios: aqueles sem envolvimento (AA2004 ou apenas AD2004) x aqueles com envolvimento (entraram em 2003 ou anos anteriores). Esse passo finalizou a etapa em relato.

Juntas, essas duas etapas (Passo V e Passo VI) permitiram a análise do atingimento dos objetivos da AA, como um dos usos instrumentais possíveis para uma política de avaliação. A próxima etapa, descrita a seguir, concentrou-se nos itens de verificação para a categoria Uso Conceitual.

3.4.8 Passo VIII: Levantamento dos itens de verificação para o Uso Conceitual da AA

O Uso Conceitual é aquele que viabiliza um entendimento generalizado sobre o objeto da avaliação, ainda que não esteja relacionado a uma tomada de decisão ou a uma ação imediata. Defende-se, no entanto, que esse uso tenha um efeito sobre a escola, ainda que não tão direto como aquele do uso instrumental. Por essa razão, além da narrativa do uso feito pela escola, buscou-se um cruzamento entre este relato e o ano de envolvimento de seus municípios na AA (Gamma) e uma comparação das suas médias de desempenho dos alunos da 4ª série em Português e em Matemática na AD 2004 (ANOVA).

No presente estudo, o indicador Uso Conceitual foi construído com três dimensões (Político-persuasório, Motivacional e Partilha) e cinco itens de verificação (os três primeiros ligados à dimensão Uso Conceitual Político-persuasório). A próxima figura ilustra a lógica dos cruzamentos feitos.

---

58 BAHIA/SEC. Avaliação de Desempenho 2004: resultados gerais e análises pedagógicas (2005).

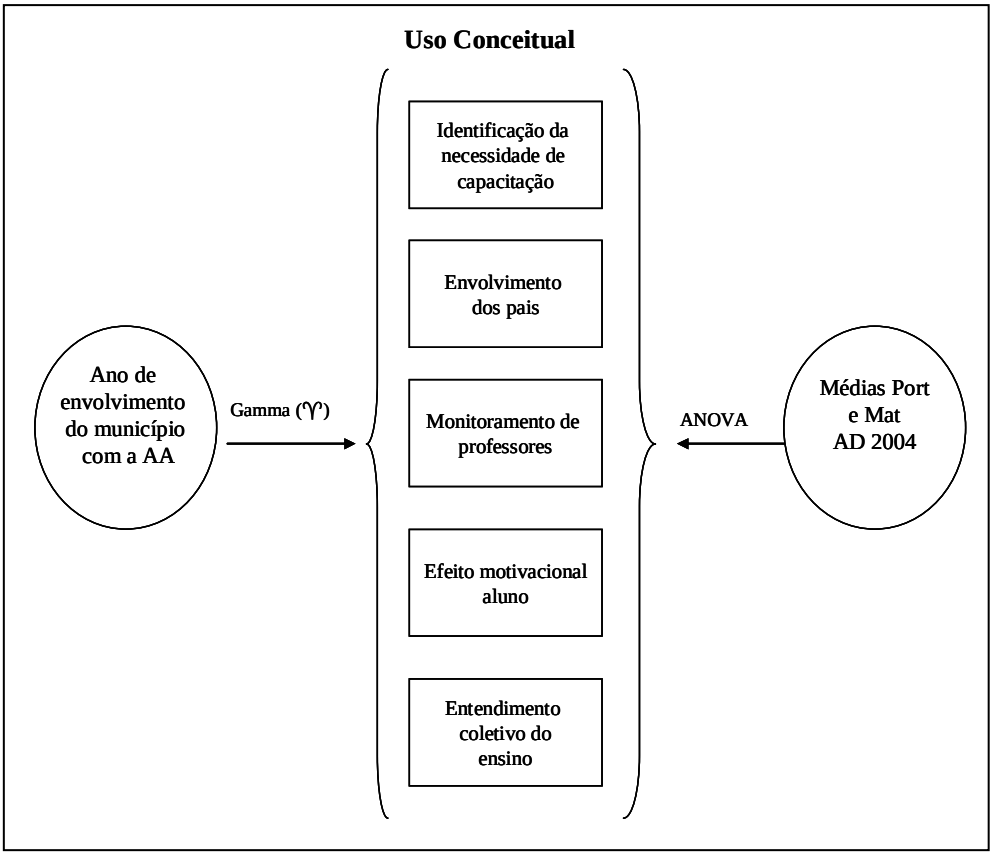


Ilustração 13: Lógica da análise dos itens de verificação do Uso Conceitual

Como posto na Fundamentação Teórica, o levantamento dos itens de verificação do Uso Conceitual é feito a partir dos relatos de *stakeholders* e usuários, diretos e indiretos, da avaliação. No presente estudo, optou-se por concentrar os esforços no levantamento dos usos junto ao *stakeholder* /usuário principal da AA, a escola pública. Novamente foi feita uma leitura nos RD para identificação das questões que tivessem perguntado às escolas, diretamente, sobre análises feitas por elas que, ainda que não necessariamente tivessem implicado ações subseqüentes, as tivessem tornado mais críticas ou que as tivessem levado a exercer algum tipo de pressão sobre suas redes ou suas equipes; sobre o efeito motivacional da AA nos alunos; ou sobre o compartilhamento de visões e idéias sobre a realidade do ensino na escola.

Após a leitura inicial dos RDs, optou-se por concentrar o estudo nos relatos das escolas encaminhados à equipe central da avaliação na AA 2004 – 3ª unidade. A base de dados já havia sido preparada para o passo IV. O quadro a seguir relaciona as questões do RD 2004 – 3ª unidade utilizadas para análise do uso conceitual, nas suas três dimensões.

Dimensão	Itens de verificação	Questões do RD 2004 – 3ª unidade	Abordagens
Político-persuasório	Identificação de capacitação	Q.16	Descrição das freqüências; comparação dos comportamentos dos grupos de escolas contrastados pelo envolvimento do município com a AA; comparação dos grupos de escolas contrastados pelas médias dos alunos da 4ª série em Português e em Matemática AD 2004.
	Envolvimento de pais	Q. 23 h	
	Monitoramento de professores	Q.22ª e Q 22 b	
Motivacional	Efeito motivacional para os alunos	Q. 18b, Q.19b, Q 20b, Q. 21b	
Partilha	Discussão e entendimento coletivos das questões do ensino	Q. 12, Q. 23ª, Q. 23b, Q. 23 f e Q 23 g	

Quadro 11: Quadro Operacional para a categoria Uso

A seguir são transcritas as perguntas do RD 2004 – 3ª unidade que deram origem aos dados utilizados nesse passo metodológico. O item de verificação Identificação de capacitação foi analisado a partir das respostas à pergunta Q. 16, transcrita a seguir:

16. A partir do diagnóstico obtido durante o ano letivo, sua escola detecta a necessidade de capacitação para os professores? Em que área(s)?

.....

.....

.....

.....

.....

.....

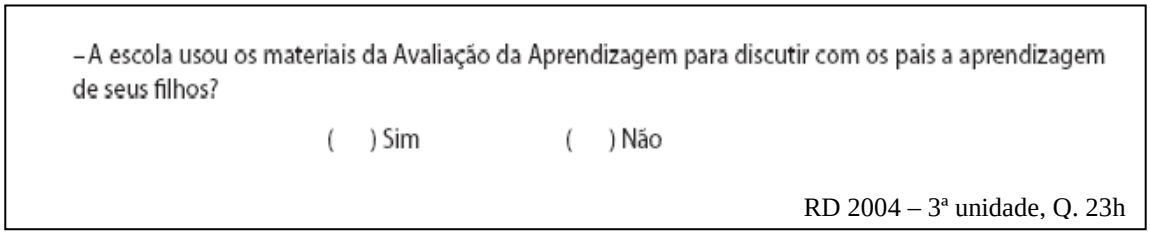
.....

Caso o espaço seja insuficiente, anexe uma página ao relatório e continue.

RD 2004 – 3ª unidade, Q.16

Os dados foram categorizados e tabulados pela equipe de avaliação de acordo com as informações enviadas pelas escolas. Nessa categorização, tanto estiveram presentes os aspectos mais gerais (necessidade de capacitação em Português) quanto mais específicos (necessidade de capacitação em produção textual, por exemplo). Por essa razão, foram criadas duas novas variáveis: capacitação em português e capacitação em matemática. Foi apenas considerado 0 - não relata e 1 – identifica necessidade de capacitação na área. Os demais aspectos de identificação de capacitação foram descartados.

Para a análise do item de verificação Envolvimento dos pais, foi utilizada a questão Q.23h, transcrita a seguir:



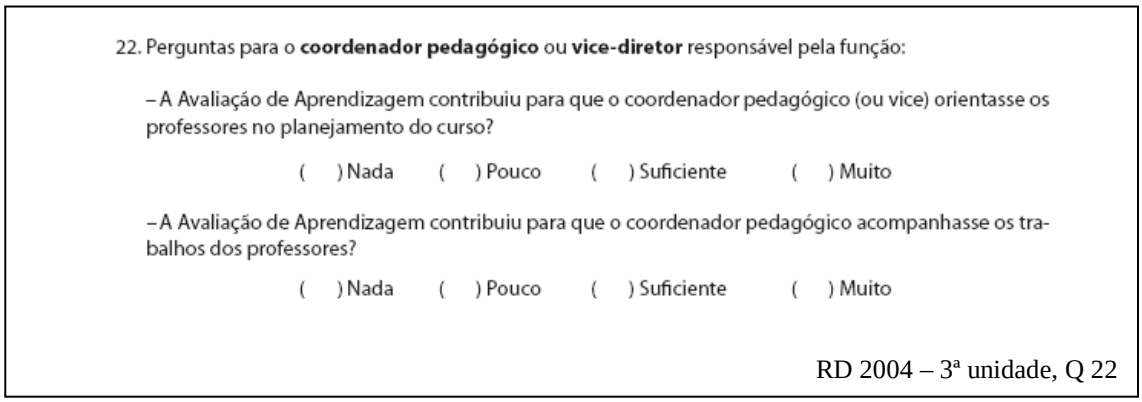
– A escola usou os materiais da Avaliação da Aprendizagem para discutir com os pais a aprendizagem de seus filhos?

( ) Sim ( ) Não

RD 2004 – 3ª unidade, Q. 23h

As respostas a essa pergunta foram tabuladas como Não relata (0), Não (1) e Sim (2). Para análise da variância, foram criados apenas dois grupos: Não (0) e Sim (1). As respostas Não relata (22 em 915) foram tratadas como *missing*.

O terceiro item de verificação para o uso conceitual político-persuasório voltou-se para o monitoramento dos professores. Foram utilizadas duas questões do RD 2004 – 3ª unidade para o levantamento de dados relativos a esse monitoramento, como pode ser visto na caixa de texto a seguir.



22. Perguntas para o **coordenador pedagógico** ou **vice-diretor** responsável pela função:

– A Avaliação de Aprendizagem contribuiu para que o coordenador pedagógico (ou vice) orientasse os professores no planejamento do curso?

( ) Nada ( ) Pouco ( ) Suficiente ( ) Muito

– A Avaliação de Aprendizagem contribuiu para que o coordenador pedagógico acompanhasse os trabalhos dos professores?

( ) Nada ( ) Pouco ( ) Suficiente ( ) Muito

RD 2004 – 3ª unidade, Q 22

Novamente, foram levantadas as freqüências simples das respostas obtidas a essas duas perguntas e, em seguida, buscou-se uma tendência de comportamento em termos do ano de envolvimento do município com a AA. Isso feito, as duas variáveis foram recategorizadas: as respostas nada e pouco foram transformadas em Pouco ou nada (0) e as respostas Suficiente e Muito foram transformadas em Suficiente e Muito (1). Dessa maneira, criaram-se dois grupos, um de percepção nula ou negativa da utilização da AA para o monitoramento e outro, de percepção positiva. Essa nova categorização permitiu uma análise de variância quando observados as médias dos resultados dos alunos em Português e Matemática na AD, por escola, em 2004.

Para análise do efeito motivacional para os alunos, foi feito o levantamento das respostas à pergunta transcrita a seguir, feita sobre a 1ª série. Tal questão foi repetida para as três outras séries do Fundamental Menor.

> – A equipe escolar percebeu um aumento no interesse em aprender, demonstrado pelos alunos de 1ª série, como efeito do trabalho com a Avaliação de Aprendizagem?
>
> ( ) Sim ( ) Não ( ) Não é possível afirmar
>
> RD2004-3ª unidade, Q.18 b

Em seguida, foi feita uma comparação entre as médias em *theta* em Português e Matemática – AD 2004 entre o grupo que respondeu sim e o grupo que respondeu não ou não é possível afirmar. Para a comparação, foi utilizada ANOVA.

A última dimensão de Uso Conceitual analisada no estudo em tela - Partilha - foi verificada por cinco questões do RD 2004 – 3ª unidade. A primeira levantou a ocorrência, na escola, das reuniões dos professores com a direção para discussão dos resultados da aplicação da AA na 3ª unidade. As demais perguntaram à escola se a AA havia contribuído para uma reflexão sobre as dificuldades encontradas pelos alunos; se houve o estabelecimento da relação dificuldade do aluno x planos de aula e prática; se houve alteração na freqüência de reuniões para discussão dos diagnósticos; e, por fim, se a busca por soluções passou a ser mais participativa. As questões e alternativas de resposta estão transcritas a seguir.

> 12. Nessa unidade, houve reunião da direção com os professores para discussão dos resultados (*Quadros-diagnóstico*)?
>
> ( ) Sim ( ) Não
>
> RD 2004 –3ª unidade

> 23. Perguntas para o conjunto de professores de 1ª a 4ª séries, dirigentes e coordenadores (ou vice-diretores que exerçam essa função):
>
> – A Avaliação de Aprendizagem contribuiu para que a equipe escolar refletisse sobre as dificuldades de seus alunos, a partir do diagnostico obtido nas unidades?
>
> ( ) Nada ( ) Pouco ( ) Suficiente ( ) Muito
>
> – A Avaliação de Aprendizagem contribuiu para que a equipe escolar relacionasse os resultados alcançados pelos alunos com seus planos de aula e com a sua prática?
>
> ( ) Nada ( ) Pouco ( ) Suficiente ( ) Muito
>
> RD 2004 – 3ª unidade, Q. 23ª e Q. 23b

– Em relação à freqüência das reuniões realizadas entre a direção e o corpo docente, para discussão dos diagnósticos dos alunos, pode-se dizer que, depois da Avaliação de Aprendizagem:

( ) Diminui ( ) Não foi alterada ( ) Aumentou

– A partir da Avaliação de Aprendizagem, a busca por soluções para se trabalhar as dificuldades dos alunos passou a ser mais participativa em sua escola?

( ) Sim ( ) Não

RD 2004 – 3ª unidade, Q. 23 f e Q. 23 g

O mesmo procedimento utilizado anteriormente, nos demais itens verificadores, foi repetido para essas cinco questões. Em primeiro lugar, foi buscada a freqüência de respostas. Em seguida, as respostas foram cruzadas com o tempo de envolvimento do município com a AA, sempre na busca por tendências de comportamento que pudessem estar associadas à AA. Por fim, as variáveis foram recodificadas de modo a que fossem obtidos dois grupos (a AA contribuiu para o compartilhamento x a AA não contribuiu) e suas médias na AD 2004 fossem comparadas.

Finalizado o levantamento do Uso Conceitual, foi feita uma análise geral do Uso feito da AA, a partir dos itens verificadores escolhidos. Em especial, buscou-se verificar a hipótese: em políticas de avaliação educacional em larga escala, os resultados são elementos pouco utilizados e é o acontecimento da avaliação que afeta as instituições em nível micro (escolas).

Os últimos passos do presente estudo foram dedicados à elaboração e revisão do texto da tese.

4. Resultados: as contribuições da política de Avaliação de Aprendizagem

Os resultados da investigação estão apresentados em duas subseções: as contribuições da política de Avaliação da Aprendizagem (AA) observadas na ótica da categoria Utilidade e, em seguida, detalhadas pela categoria Uso.

Análise da política de Avaliação de Aprendizagem a partir da categoria Utilidade

Nessa subseção, os resultados da análise da AA estão dispostos em nove segmentos: os sete primeiros são relativos aos indicadores da Categoria Utilidade, adaptados do *checklist* de Stufflebeam para os padrões do JCSEE, o oitavo faz uma síntese da AA perante esses sete indicadores e o nono segmento detalha a percepção de utilidade sobre a AA relatada pelos *stakeholders*/usuários principais: as escolas públicas.

Cada uma das sete primeiras subseções é iniciada com um texto que descreve o comportamento da AA pelos itens verificadores do *checklist* (U1 a U7) para que, ao final, seja apresentado um quadro síntese. Já a subseção que trata do U8 apresenta inicialmente as respostas de 290 escolas em uma consulta, feita pelo RD 2003 – 3ª unidade, sobre os efeitos da AA em 20 áreas previamente selecionadas e, em seguida, aprofunda a noção do efeito da AA pelo posicionamento das escolas no RD 2004 – 3ª unidade.

4.1.1 Análise do U1: Identificação dos *stakeholders* da política de Avaliação de Aprendizagem

Como detalhado na Subseção 2.4.2.1, o primeiro indicador da categoria Utilidade lida com a identificação de *stakeholders* e com o levantamento e o atendimento às suas necessidades, especialmente no delineamento da avaliação.

Para a apresentação do conjunto de *stakeholders* da AA, optou-se pela utilização do formato do Quadro 5 (Subseção 2.4.2.1), que sintetiza as instâncias envolvidas pelas políticas de avaliação, seu nível de implementação (se político, técnico central, técnico não central e não técnico), acrescido do grau de prioridade. A identificação dos *stakeholders* da AA foi feita por meio da

leitura dos documentos encaminhados às escolas, à SEC e às prefeituras municipais durante o primeiro ciclo da política[59].

Política de AA	Nível	Grau de Prioridade	Identificação	Grau de atendimento
Cliente	Político	Alto	SPDE na SEC de 1999 a 2002; SUPAV na SEC em 2003-2004.	Médio
Stakeholders	Político	Alto	Secretário de educação do Estado	Médio
	Político	Alto	Secretários de educação dos municípios envolvidos	Baixo
	Técnico central, com atuação local	Médio	Líderes dos demais projetos prioritários do programa de reforma do Governo	Médio
	Político/técnico central	Baixo	Dirigentes e técnicos dos demais departamentos e setores do órgão central	Baixo
	Técnico não central	Baixo	Representações e técnicos das DIREC	Baixo
	Não técnico	Baixo	Alunos e pais de alunos	Baixo
Stakeholders /Usuários diretos	Técnico não central	Alto	Dirigente escolar	Média
	Técnico não central	Alto	Professores nas escolas	Médio
	Técnico não central	Alto	Coordenador pedagógico nas escolas	Médio
	Técnico não central	Médio	Técnicos das coordenações pedagógicas dos municípios envolvidos	Baixo
Usuário	Não direto	Baixo	Academia	Média

Quadro 12: Panorama dos principais *stakeholders* e usuários da política de Avaliação de Aprendizagem e dos seus respectivos graus de prioridade, nível de atuação e grau de atendimento quanto às expectativas originais.

De acordo com o disposto nos documentos oficiais, até o início de 2003, havia uma clara identificação do cliente da AA, a SPDE, responsável pelo acompanhamento do convênio 444/99, estabelecido entre SEC, UFBA e FAPEX para sua implementação. Com a mudança de governo em 2003, a SUPAV assumiu a avaliação com alguma dificuldade, já que seus novos integrantes não haviam participado anteriormente dos processos de negociação da política. Até o final daquele ano, foram mantidas várias reuniões para ajustes de percepção e discussão de novas demandas do órgão central, especialmente em relação à avaliação da alfabetização, que não

59 Para a relação integral dos documentos consultados, ver o capítulo Metodologia, Subseção 3.3.

chegou a ser atendido, pela descontinuidade do Projeto, ainda que estudos preliminares tivessem sido conduzidos.

Desde a concepção inicial, foi estabelecido que a escola (e nela o dirigente, os professores e coordenadores pedagógicos) seria o centro da política, sendo a principal instância de utilização e uso. O delineamento negociado não mudou tal grau de prioridade. É interessante perceber que, embora todo o esforço do programa de reforma estivesse voltado para a melhoria da qualidade da educação, visando ao atendimento de alunos e, por conseguinte, de seus pais, a política de AA teve o aluno como fonte de dados, não sendo um alvo principal para o uso direto da avaliação. Por essa razão, no quadro anterior, aos alunos e seus pais foi conferido um grau de prioridade baixa.

O grau baixo de atendimento às secretarias municipais é decorrente do fato de que as escolas rurais não foram incluídas na avaliação. Em vários casos, as escolas urbanas formavam um percentual pequeno da rede municipal de educação e a AA, aplicada apenas a elas, teve implementação muito restrita. O caso mais interessante foi Nilo Peçanha. Na área urbana, não havia escolas de 1ª a 3ª séries. De acordo com o Relatório de Conclusão do Convênio 444/99, em 2001, a AA envolveu os alunos das 1ª e 2ª séries do Ensino Fundamental e ciclo básico I de 1.532 escolas públicas urbanas de 126 municípios. Em 2002, a AA foi expandida e 2.700 escolas de 273 cidades receberam testes de 1ª a 4ª séries nas três primeiras unidades do ano letivo. Após ajustes, a AA abrangeu 2.610 escolas em 271 municípios. A mesma clientela foi atendida em 2003, apenas com a não participação de um município. Por fim, na segunda unidade de 2004, a avaliação alcançou 299 municípios, como mostra a Tabela 02 a seguir.

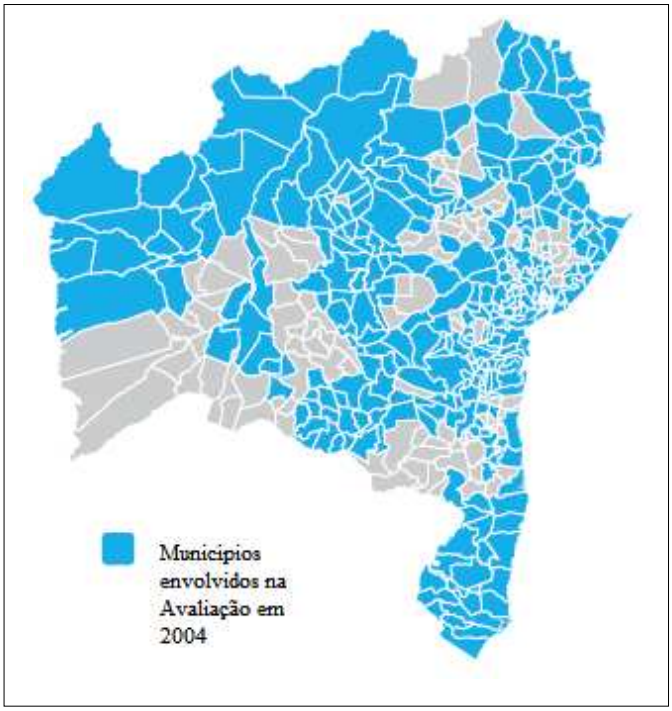
Tabela 2: Panorama de Expansão da AA de 2001 a 2004.

Envolvimento com a AA	**Ano**			
	2001	**2002**	**2003**	**2004***
DIREC	31	31	31	33
Municípios parceiros	126	273	272	272/299
Escolas estaduais	sem registro	949	929	784/862/759
Escolas municipais	sem registro	1.751	1.727	1.783/1.989/2.087
Total de escolas	1.532	2.700	2.656	2.567/2.851/2.846
Alunos (estimativa)	245.000	700.000	600.000	311.613/644.836
			*Dados das três unidades letivas	

Fonte: Relatório de Conclusão do Convênio 444/9 e Relatório Síntese de Monitoramento 2004-3ª unidade.

É essa expansão ao longo dos anos que permitiu, na presente investigação, a definição do envolvimento das escolas com a AA a partir da assinatura da parceria dos municípios com o

Governo do Estado da Bahia. Optou-se, como visto na Metodologia, pela composição de dois grandes grupos: um deles abrange as escolas situadas em municípios que não tiveram relação com a AA ou a estabeleceram em 2004, não tendo, portanto, tempo para a consolidação da política; o segundo compreende as escolas em municípios que foram envolvidos em 2003 ou anos anteriores. O mapa a seguir mostra a distribuição dos municípios abrangidos tanto pela AA quanto apenas pela AD em 2004.



Fonte: BAHIA/SEC. Projeto de Avaliação Externa 2004

Ilustração 14: Abrangência da AD em 2004 na Bahia, que corresponde ao total de municípios envolvidos com a AD e AA naquele ano.

Os textos e documentos da AA permitem a identificação de cliente e *stakeholders* principais, mas a definição da prioridade dos demais *stakeholders* e usuários não é tão clara. Foram observadas demandas para a AA a partir do nível técnico central, do nível técnico não central, das representações municipais, e dos projetos parceiros no Programa Educar para Vencer. A definição de prioridades disposta no Quadro 12 foi feita a partir da análise das comunicações (*e-mails* para a coordenação do Projeto de Avaliação, ofícios ou relatos em documentos oficiais) entre a AA e tais *stakeholders*.

Por fim, ao discutir os *stakeholder*s e usuários, é interessante registrar que, a partir da composição da equipe técnica da avaliação sob coordenação do ISP[60] em meados de 2000, houve um cuidado com os usuários indiretos (membros da Academia). Tal cuidado foi concretizado, por exemplo, pelo incentivo ao uso de dados (como o PIPEP 2002)[61] ou pela preparação das bases de dados de modo a que pudessem ser disponibilizadas para estudos[62].

Analisados os itens de verificação quanto à identificação de *stakeholders*, os demais itens do indicador U1 voltam-se para a determinação de suas demandas e para o seu atendimento, também sintetizados no Quadro 12 anterior. No final de 1999 e no início de 2000, após a aplicação de provas da AD pela Fundação Carlos Chagas, a SEC (SPDE e Secretário) definiu um curso de ação que resultou na parceria com a UFBA e na contratação de uma agência consolidada no campo avaliativo para transferir tecnologia para a Bahia. O projeto original, inviável em termos de tempo, custo e equipe, foi redesenhado até o formato contratado, de alguma maneira passando pelas questões já discutidas na Subseção 2.2.4, que tratou do delineamento de uma experiência de avaliação. Contribuiu para esse redesenho a entrada do ISP/UFBA na negociação, em março de 2000, com a discussão do plano de trabalho para o termo aditivo II ao convênio 444/99. O redesenho alterou uma série de características do projeto original e reduziu a amplitude de atendimento aos *stakeholders*. Durante a negociação inicial do desenho da política de Avaliação, não há registro de consulta à escola, às representações municipais ou mesmo às outras coordenações na SEC.

Naturalmente, em decorrência de adequações na implementação do Programa Educar para Vencer e do Projeto Bahia, as demandas dos atores também foram sendo alteradas ou novas demandas foram colocadas nos quatro anos do ciclo da AA. Algumas situações ilustram esse comportamento. É emblemático, por exemplo, que a avaliação de impacto do Projeto Bahia, em 2002, tivesse demandado informações da AA ao nível da escola, quando, no redesenho em 2000, havia sido decidido um monitoramento por amostra, sendo impossível o fornecimento de tais dados. Já para algumas escolas, provas padronizadas aplicadas em freqüência bimensal feriram as

---

60 O convênio 444/99 foi firmado com a UFBA sem identificação de um departamento ou unidade responsável por sua condução. O reitor solicitou ao Prof. Robert E. Verhine, então diretor do ISP, que fizesse uma análise das unidades da UFBA capazes e interessadas no trabalho. Ao final, o próprio ISP assumiu a implementação da política, o que valeu um aditivo ao convênio original e grandes modificações no plano de trabalho. Vale ressaltar que a primeira unidade da UFBA consultada foi a FACED, em outubro de 1999. A mesma se negou a assumir o Projeto por entendê-lo um "pacote fechado".

61 PIPEP – Programa de Incentivo aos Pesquisadores do Ensino Público. Em sua única edição, foi vencedor Cláudio Pondé Avena, com o trabalho Determinantes do Rendimento Escolar: Evidências a partir da avaliação externa da aprendizagem do ensino fundamental da rede pública do Estado da Bahia.

62 Desde o início, o Projeto de Avaliação Externa adotou uma política de incentivo ao uso de suas informações pela comunidade científica em paralelo ao compromisso com o sigilo sobre o desempenho das unidades escolares.

crenças do corpo docente, imerso em abordagens ditas construtivistas. Esse embate ocorreu também na própria SEC, em relação à SUPEN, que oferecia, especialmente no início, severas críticas ao desenho avaliativo adotado pela SPDE (em especial, contra a imposição de ritmo).

Além disso, os objetivos do Projeto de Avaliação Externa, de maneira geral, não englobaram o monitoramento e o levantamento do impacto de outras ações da SEC ou o uso da avaliação, especialmente a AA, como ferramenta de prestação de contas. A delimitação dos objetivos da Avaliação Externa, nas suas duas vertentes, também não atendeu inteiramente aos demais projetos prioritários, parceiros no desenvolvimento da reforma proposta pelo governo. A frustração dessas expectativas contribuiu para o surgimento de problemas de articulação, com impacto para a percepção de utilidade e para a concretização dos usos da avaliação por estes *stakeholders*. São exemplos das diferenças entre as expectativas dos outros projetos e dos parceiros municipais do programa de reforma e seu (não) atendimento:

1) O Projeto de Regularização de Fluxo de 1ª a 4ª mencionou em seu *folder* de divulgação "a confirmação" do seu êxito como sendo um produto da Avaliação Externa e não foi atendido. As atividades da AA, ainda que voltadas para a clientela de 1ª a 4ª, não incluíram as turmas de Fluxo. A demanda por avaliação do Fluxo e do Ciclo Básico também foi apresentada por várias escolas, mas não foi possível um delineamento específico para atendê-las.

2) O Projeto de Fortalecimento da Gestão Escolar teve como expectativa a obtenção dos resultados escolares de maneira a poder identificar melhores e piores experiências, bem como a comparabilidade de dados por escola ao longo dos anos. A AA não permitia esse tipo de identificação e a comparabilidade era dada pela AD.

3) Os *stakeholders* vinculados de alguma forma às redes municipais (secretários, equipe técnica das secretarias, líderes do Projeto de Capacitação da Gestão Municipal) foram, desde o início, apenas parcialmente atendidos pela AA vez que, para muitos municípios, boa parte das escolas estava situada na zona rural, não compreendida pela aplicação dos testes e instrumentos. Além disso, com a implementação da AA, secretários municipais e algumas representações da Secretaria Estadual apresentaram, ao longo do tempo, uma demanda específica: dados individualizados das escolas de modo a poder regulá-las ou controlá-las. No caso dos secretários municipais, a necessidade do controle era justificada especialmente nos casos nos quais os dirigentes escolares eram uma escolha direta do prefeito e tendiam, nesse contexto, a não se reportar concretamente às secretarias de educação. Já para alguns setores da SEC, os resultados eram demandados por uma reação importante ao processo de

autonomia: muitos técnicos jamais acreditaram que a maior parte da rede pública pudesse funcionar sem tutoria constante do órgão central.

Quanto ao atendimento das demandas da escola, durante a implementação da AA, percebe-se[63] que várias modificações foram feitas ao longo do seu primeiro ciclo com esse objetivo. As mudanças, relatadas nos documentos de conclusão da AA a cada ano, foram feitas no sentido de incluir a avaliação da produção textual, de inserir o professor no diálogo direto com a equipe central da avaliação (inclusive com a criação de um guia diagnóstico), de ajustes logísticos visando a entrega de material no tempo e na quantidade certas. Muitos dos pedidos feitos (e não atendidos) referiam-se à adequação das provas "à realidade do município", à correção feita pela equipe central da Avaliação ou à diminuição da freqüência de aplicação de provas, reduzindo assim a carga de trabalho para docentes e coordenadores já tão atribulados. Especialmente para escolas municipais, houve uma demanda de maior articulação entre a proposta da AA e a linha de trabalho adotada pela secretaria da educação municipal.

Por fim, a leitura dos Relatórios Síntese dos RD, produzidos ao longo das três unidades dos quatro anos de implementação da AA, anuncia uma baixa capacidade instalada em muitas escolas, o que as fizeram demandar formação e capacitação dos docentes e soluções prontas para os problemas que enfrentavam. Em termos desse atendimento, o delineamento da AA não considerou intervir diretamente na capacitação dos docentes ou na melhoria de infra-estrutura das escolas (inclusive quanto a recursos humanos), exceto pelo envio de materiais didáticos (vídeos e seus manuais) como ferramenta de remediação. Para muitas unidades escolares, essa ação não foi suficiente. Como discutido anteriormente, é comum uma expectativa para a avaliação que ultrapasse as ações avaliativas e se inscreva nas intervenções ou na própria gestão. Na Bahia não foi diferente.

Sintetizando a análise do comportamento da AA quanto aos itens de verificação do indicador U1, pode-se ressaltar: a identificação do cliente e do grupo principal de *stakeholders* (escola), mas um senso difuso dos demais; a não inclusão da maior parte dos *stakeholders* no processo de negociação da política; o levantamento das demandas dos *stakeholders* feito após a implantação da AA e o atendimento parcial às solicitações das escolas; a frustração de expectativas para a avaliação dos demais projetos parceiros no programa de reforma; a baixa articulação Estado – Município, com algum conflito para aceitação da AA em alguns deles, especialmente aqueles que adotavam uma linha construtivista; a resistência interna no órgão central ao delineamento da AA;

63 Ver os folhetos Avaliação de Aprendizagem: Participação e Contribuições das Escolas em 2002 e Avaliação de Aprendizagem: Participação e Contribuições das Escolas em 2003.

as demandas de escolas, representações regionais, representações municipais e corpo técnico da SEC canalizadas para a AA quando, em verdade, não faziam parte de uma ação de avaliação externa (como as demandas por capacitação docente).

O U1 é um indicador de qualidade da avaliação que permite a observação da identificação dos *stakeholders* e do atendimento de suas demandas. A análise da AA à luz do U1 aponta para uma utilidade baixa para alguns dos segmentos envolvidos pela política e para uma utilidade média para as escolas que, tendo recebido o "pacote" da avaliação, o viram modificar-se para atendê-las. Dos dez itens verificadores para esse indicador, apenas cinco podem ser marcados positivamente, enquanto os outros ou não foram observados ou o foram de maneira muito restrita.

Comportamento da AA quanto aos itens de verificação do U1

- ✓ Identifica de maneira clara o cliente da avaliação.
- ✓ Envolve as lideranças na identificação de outros *stakeholder*s.
  Levanta as necessidades de informação dos *stakeholder*s potenciais.
  Utiliza *stakeholder*s para identificação de outros *stakeholder*s.
- ✓ Com o cliente, classifica os *stakeholder*s pela sua importância relativa.
- ✓ Envolve os *stakeholder*s durante o processo avaliativo (alguns deles, embora nem sempre os atenda).
  Mantém aberta a avaliação para novos *stakeholder*s.
  Atende às necessidades avaliativas dos *stakeholder*s.
- ✓ Atende uma gama adequada de *stakeholder*s individuais (quando consideradas as escolas – nível micro).
  Atende uma gama adequada de organizações (as secretarias municipais foram pouco atendidas).

4.1.2 Análise do U2 – Credibilidade do avaliador na Avaliação de Aprendizagem

Os itens de verificação do indicador Credibilidade tratam da competência do avaliador ou da equipe avaliadora e do comportamento desses indivíduos. A SEC, ainda em 1999, percebeu que a questão de credibilidade teria um peso significativo para suas avaliações, especialmente porque seu programa de reforma envolvia a avaliação das redes municipais que poderiam, em algum momento, sentir-se em desvantagem em comparação com a rede estadual, cliente da AA. Cinco dos seis projetos do Programa Educar para Vencer foram conduzidos pela Fundação Luis Eduardo Magalhães. O sexto projeto – Avaliação - foi entregue à UFBA por três razões: isenção político-partidária reconhecida em termos de condução de pesquisa; competência instalada (ainda que não em avaliação em larga escala) e, como ente federal, condição externa ao sistema público estadual de Educação, o que assegurava o caráter externo da avaliação. Esse movimento de aproximação das universidades para a condução de avaliação externa era relativamente comum no Brasil à época (ver a USP e a avaliação em São Paulo e a UFJF e a avaliação em Minas Gerais, por exemplo). Não houve grande questionamento quanto a essa escolha por parte dos *stakeholders* na

SEC ou nas secretarias municipais. Como discutido no marco teórico (Subseção 2.4.2.2), a falta de capacidade para analisar o trabalho de avaliação é, muitas vezes, compensada pela percepção de integridade de quem a conduz.

Em termos da burocracia média central, o estranhamento esperado com a academia foi menor por duas razões: os pesquisadores não se envolveram diretamente com a implementação da política, por um lado, e os instrumentos foram desenvolvidos por professores das redes estaduais e municipais. Para muitas das escolas, entretanto, os teóricos da universidade (de modo geral) não conheciam sua realidade e, portanto, não saberiam criar uma avaliação adequada. Esse posicionamento foi apresentado em diversas reuniões de entrega de resultados da AD, de capacitação de professores para elaboração de itens, ou em respostas nos RDs.

Essa percepção das escolas foi, de certo modo, agravada pela escolha de uma empresa americana para transferência de tecnologia em avaliação para a equipe central do Projeto entre 2000 e 2002. Em 1999, não havia, na Bahia (como de resto no Nordeste), profissionais com experiência no delineamento e na condução de avaliações em larga escala. Até então, mesmo o Ceará (pioneiro na avaliação própria) utilizava itens produzidos pela equipe do INEP e contava com professores da Universidade Federal do Ceará para seu tratamento e análise. Os estados do Sudeste e o Distrito Federal / Brasília centralizavam esses profissionais, já comprometidos com outras avaliações. Nesse cenário, a Bahia optou por formar a competência em avaliação. De acordo com informações disponíveis no Relatório de Conclusão do Convênio 444/99, foi feita uma seleção de agências de avaliação que tivessem interesse em assumir a construção das avaliações em 2001 e 2002 e, ao fazê-lo, transferir tecnologia para a equipe local. Das agências contatadas (duas no Brasil e três no exterior), apenas a *American Institutes for Research* (AIR) demonstrou interesse. Essa empresa, como primeira tarefa, delineou o perfil da equipe local. Novamente, de maneira geral, não houve questionamento quanto à credibilidade da empresa frente a um sólido currículo em pesquisa social e avaliação[64]. Para as escolas e alguns setores da SEC, no entanto, a dúvida manifesta em princípio era relacionada à capacidade de uma empresa de língua inglesa escrever testes *em* português e, mais grave ainda, construir testes para avaliar alunos em questões *sobre* Língua Portuguesa. Esse problema foi amenizado pela ampla divulgação de que os itens de testes seriam escritos por professores da rede pública na Bahia e, posteriormente, pela conclusão do contrato da AIR em 2002.

64 Em seu *portfolio*, a AIR relacionava clientes do porte e demanda do Pentágono americano. Seu representante local em Salvador era um profissional aposentado de Harvard. Esse tipo de informação era utilizado como "símbolo" de competência e funcionava em termos de credibilidade.

Diante dos prazos e dos produtos do delineamento avaliativo, SEC e UFBA concordaram em contratar profissionais (de fora do setor público) com experiência em áreas correlatas e em formá-los no campo da avaliação. A seleção desses profissionais incluiu o estabelecimento de compromisso para dedicação integral ao Projeto de Avaliação Externa e uma avaliação feita por uma banca composta por professores seniores da UFBA (inclusive com entrevista em inglês). Durante o primeiro ciclo da AA, a equipe foi capacitada, inicialmente com o apoio da AIR e, continuamente, pela participação em oficinas, seminários e cursos oferecidos na Bahia, no Brasil e, excepcionalmente, nos EUA. Dentre os aspectos de capacitação, estavam aqueles voltados para o respeito e a atenção a questões de gênero, *status* sócio-econômico, raça, e diferenças culturais e de linguagem. Disso resultou o Manual de Revisão de Viés, usado como base para as oficinas de elaboração de itens.

Além dessas estratégias com objetivo de assegurar credibilidade às suas políticas de avaliação, SEC e UFBA constituíram um Conselho Consultivo que, durante os três primeiros anos do Projeto de Avaliação, reuniu-se para acompanhar as ações e fazer correções de rumo, quando e se necessárias. O Conselho foi constituído por: Superintendente SPDE-SEC; Diretor do ISP/UFBA e responsável pela Avaliação; Representante da AIR-Washington; Representante da AIR-Brasil[65]; Consultor-idealizador do Programa Educar para Vencer; e Coordenador do Projeto de Avaliação Externa. As reuniões do Conselho favoreceram o atendimento aos itens de verificação "Mantém-se a par das forças políticas e sociais" e "Mantém as partes interessadas informadas sobre o progresso da avaliação". Os demais *stakeholders* foram informados por meio de folhetos (2002 e 2003) e de relatórios síntese (2004), a partir do tratamento de dados oriundos do RD, a cada unidade letiva, e da análise dos resultados dos testes da amostra controlada.

Uma outra característica da AA recebeu grande atenção no início da implementação por questões de credibilidade: o fato de as escolas aplicarem e corrigirem seus próprios testes poderia conduzir à obtenção de informações ruins por mau uso dos instrumentos, por desconhecimento dos processos de avaliação, ou ainda por má fé. A AA enfrentou o seguinte ponto: era essencial que o diagnóstico da escola fosse disponibilizado no início da unidade letiva seguinte à aplicação do teste, para permitir ajustes no planejamento de curso e de aulas. Nas condições que se apresentavam em 2001, era impossível à equipe central da avaliação fazer a correção dos testes,

65 A presença de Simon Schwartzman nesse Conselho (como representante da AIR Brasil) garantiu, entre outras questões, que fosse possível um levantamento sócio-econômico-educacional dos alunos da AD, com enriquecimento

tratar os dados, elaborar e divulgar relatórios individualizados em menos de quinze dias. Tal exigência era ainda mais difícil de atender porque havia previsão de um período de aplicação, mas não havia uma data fixa: a aplicação dos testes era condicionada à finalização da unidade letiva e nem sempre os calendários municipais coincidiam com o calendário da rede estadual. Nesse panorama, houve um esforço de comunicação às escolas sobre a necessidade de cumprimento de um padrão de aplicação, correção e diagnóstico.

Além disso, era importante a construção da cultura de avaliação no Estado, com envolvimento de diretores e professores no desenrolar da política. Para ajudar no processo de consolidação dessa cultura, o delineamento avaliativo usado recebeu caráter *low stakes*: os resultados de cada escola estariam restritos a elas mesmas. Esperava-se, com isso, que as escolas não se sentissem pressionadas pela AA e que não utilizassem estratégias para (má) manipulação de resultados. As questões de credibilidade foram, nesse assunto, endereçadas. A decisão pelo caráter *low stakes*, no entanto, teve várias conseqüências. Uma das hipóteses para o baixo encaminhamento dos RDs à equipe central (por volta de 50%) é atrelada a ele. A outra foi o não endereçamento de uma demanda freqüente da esfera municipal: receber os resultados de suas escolas.

Analisadas as questões de credibilidade diretamente associadas à equipe avaliadora e à condução da avaliação, os itens de verificação do U2 que implicam a interação equipe de avaliação – *stakeholders* para ajudar esses últimos a entender o plano avaliativo (e seus aspectos de qualidade técnica e operacional) e que tratam da resposta às críticas e sugestões foram analisados pela leitura da correspondência entre a coordenação da equipe central e escolas ou secretarias de educação. Diante da abrangência da AA e do tamanho da equipe central (15 pessoas diretamente ligadas ao Projeto), contatos pessoais foram realizados em fóruns e reuniões conduzidas pelos demais projetos do Educar para Vencer. Muito mais frequentemente a comunicação se deu de forma escrita ou, para o esclarecimento de dúvidas e contatos imediatos, pelo telefone. Dúvidas e críticas encaminhadas ao Projeto por meio do RD foram respondidas por ofícios. À exceção do questionamento das escolas sobre a possibilidade de pessoas da universidade entenderem sua realidade ou da empresa americana fazer testes em Português, os demais aspectos do indicador U2 apontaram para a presença de elementos que sugerem qualidade para a AA. Dos dez itens de verificação, a análise da AA verificou o atendimento de nove, como pode ser visto na síntese a seguir.

---

da base de dados mais tarde utilizada em pesquisas de mestrado e doutorado. Essa concepção foi posteriormente trazida para a AA.

Comportamento da AA quanto aos itens de verificação do indicador U2

- ✓ Emprega avaliadores competentes.
- ✓ Emprega avaliadores nos quais os *stakeholders* confiam.
- ✓ Emprega avaliadores que podem responder a preocupações dos *stakeholders*.
- ✓ Emprega avaliadores que adequadamente respondem a questões de gênero, *status* socioeconômico, raça, e diferenças culturais e de linguagem.

  Assegura que o plano de avaliação atende às principais preocupações dos *stakeholders*.
- ✓ Ajuda os *stakeholders* a entenderem o plano de avaliação.
- ✓ Fornece aos *stakeholders* informações sobre aspectos de qualidade técnica e operacional do plano de avaliação.
- ✓ Responde adequadamente às críticas e sugestões dos *stakeholders*.
- ✓ Mantém-se a par das forças políticas e sociais.
- ✓ Mantém as partes interessadas informadas sobre o progresso da avaliação.

4.1.3 Análise do U3 - Escopo e seleção da informação pela Avaliação de Aprendizagem

Como já discutido em 2.4.2.3, o indicador U3 pode ser analisado a partir de dois pontos principais: 1) a incorporação, pelo avaliador, da demanda levantada junto aos *stakeholders* (em ordem de prioridade) na definição do escopo sob avaliação, e 2) a definição da suficiência em termos de coleta de dados (em especial, a alocação de esforços avaliativos para os elementos prioritários desse escopo).

Documentos originais do Projeto de Avaliação Externa, antes da negociação do plano de trabalho com a UFBA e com a AIR, apontavam para uma avaliação que englobasse cinco disciplinas no Ensino Fundamental (Português, Matemática, Ciências, Geografia e História), da 1ª à 8ª série. Após março de 2000, ficou definido junto à SEC (SPDE e Secretário) que os esforços seriam concentrados em Língua Portuguesa e Matemática, consideradas disciplinas críticas tanto por seu impacto para a aprendizagem das demais, quanto pelo baixo desempenho da Bahia em outras avaliações, como o SAEB. O escopo definido para a AA restringiu-se à avaliação diagnóstica quanto ao desempenho de alunos (por turma) nas disciplinas de Português e Matemática e não incluiu a análise do mérito das abordagens de ensino ou de planejamento de cursos. Essa análise deveria ficar sob responsabilidade das próprias escolas e ser conduzida a partir dos resultados das turmas nas provas encaminhadas. O valor, para julgamento da proficiência das turmas nas duas disciplinas, foi determinado a partir de uma linha de corte estabelecida para cada teste aplicado.

Da leitura dos relatórios síntese dos RD, foram identificadas várias manifestações das escolas, encaminhadas à equipe central, no sentido da ampliação do escopo para inclusão de outras disciplinas. Nos fóruns e reuniões para esclarecimento sobre a avaliação, professores criticaram a redução do escopo para Português e Matemática, em especial por considerarem que a não escolha

das “outras disciplinas” por uma ação prioritária do Estado significava que o governo atribuía-lhes menor importância. Esse posicionamento remete à literatura já discutida: a simples escolha de um elemento como item de avaliação lhe confere uma saliência frente aos outros que pode ser percebida, em uma hierarquia de valores, como aquilo que é mais importante, em detrimento dos demais. Nas análises feitas por *stakeholders* e usuários, raramente são pensadas as limitações de técnica, custo ou tempo que moldam um processo avaliativo; conta o caráter simbólico.

Para as duas disciplinas sob avaliação, houve a composição de matrizes de referência organizadas por série / unidade letiva, como síntese do escopo a avaliar. A matriz, por disciplina e por série, foi organizada a partir de dimensões (por exemplo, Português para a 1ª série foi dividida em pré-leitura, leitura, leitura e escrita) e, nelas, os descritores foram distribuídos por quatro blocos de 200 horas de aulas. Dessa forma os testes encaminhados pela AA foram associados a um recorte da expectativa de currículo mínimo a ser trabalhado com os alunos.

As matrizes de referência foram distribuídas para todas as escolas, DIRECs, secretarias municipais e SEC mais de uma vez durante o ciclo da AA, como atestado pelos vários registros de processos licitatórios para a impressão e distribuição dos materiais. Para sua elaboração, de acordo com informações disponibilizadas na Matriz de Referência da Avaliação de Aprendizagem para a 1ª e 2ª séries e CBA I, em outubro de 2000, foi conduzida uma oficina de elaboração de descritores, com cerca de 70 professores das redes pública e privada da capital e do interior do Estado da Bahia[66]. Não houve entrevista de *stakeholders*, como previsto no item de verificação de Stufflebeam (1999), mas a etapa de definição do escopo foi negociada por um número muito maior de *stakeholders* que aquele envolvido na negociação anterior para o delineamento final da política da AA, apresentado na Subseção 4.1.1.

Ainda da leitura do capítulo introdutório das matrizes de referência da AA, identificam-se, como fontes para definição dos descritores: os objetivos nacionais apresentados nos PCN, os indicadores constantes no Diário de Classe do CBA I, os livros e as práticas didáticas comumente utilizadas na Bahia e, finalmente, “as experiências e sensibilidade como educadores” do grupo de professores envolvidos[67]. Dessa maneira, buscou-se evitar discrepâncias entre os parâmetros curriculares nacionais, as orientações adotadas na Bahia e as matrizes de avaliação.

66 Matriz de Referência de 1ª e 2ª séries, 2ª ed., p. 14-15
67 Em 2000, a Bahia não havia adotado um currículo mínimo que pudesse ser utilizado para tal trabalho.

Vários descritores propostos nessas oficinas foram retirados da matriz de referência, o que limitou o escopo sob avaliação. Foi o caso, por exemplo, do desenvolvimento da oralidade nas séries iniciais. A definição final dos descritores levou em conta: aplicação em larga escala, testes em formato de lápis e papel, além dos aspectos de seqüência instrucional, abrangência e pertinência de cada descritor proposto. Mais uma vez, foram endereçadas as preocupações sobre possíveis interpretações das escolas sobre a não priorização de competências importantes, já que as mesmas haviam sido excluídas das matrizes. Um esforço de esclarecimento dos *stakeholders* quanto a essa questão pode ser percebido nos diversos instrumentos de comunicação utilizados pela AA. O mesmo procedimento foi adotado posteriormente para a elaboração das matrizes de Língua Portuguesa e Matemática para as 3ª e 4ª séries, sendo que essa oficina foi realizada com a participação de 56 professores.

A proposta original de descritores para composição das matrizes, por série e por disciplina, foi submetida a um comitê de validação. O comitê foi composto por diversos representantes do nível técnico do órgão central, do técnico não central (incluindo vários professores que haviam participado das oficinas anteriores), do não técnico (pais), além de representantes de universidades. Duas questões emergem, sem que haja uma resposta, da análise do U3: o que significa envolver *stakeholders* na definição do escopo quando se trata de avaliação em larga escala? e como envolver pais em uma discussão técnica?. No universo previsto de expansão para 417 municípios e com uma atuação inicial em 126 deles, espalhados por todas as regiões da Bahia, em uma abordagem conjunta nas redes estadual e municipais, a escolha de *stakeholders* para a negociação do escopo da AA não foi compreensiva. Por mais que sejam chamadas representações sindicais, coordenações nas representações regionais e buscadas contribuições em todas as regiões, a escola pode não se sentir representada.

Esse fenômeno ocorre com o SAEB, por exemplo, quando observadas as críticas feitas pelos estados da Federação. Como visto na Subseção 2.2.4, a definição de escopo da avaliação faz parte tanto da etapa político-conceitual como da etapa técnica. Nesse sentido, as representações participantes na definição do escopo de uma avaliação precisam conhecer em profundidade o objeto a ser avaliado. Um dirigente não formado em matemática ou sem experiência no seu ensino não deveria ser convidado a participar de uma oficina de elaboração dos descritores dessa disciplina, por exemplo. Por outro lado, a participação exclusiva de especialistas na disciplina, sem uma intimidade com a sala de aula nas séries avaliadas, poderia dificultar a elaboração de descritores e, especialmente, prejudicar o processo de alocação dos descritores nas quatro unidades letivas.

A necessidade de articulação do político e do técnico sobressai-se ainda mais nos comitês de validação do escopo (negociado tecnicamente antes), porque esses comitês contam com representação da sociedade, dentre os quais pais de alunos. Ao longo do ciclo da AA, foi possível observar que pais sem conhecimento do assunto a ser avaliado tenderam a passar a reunião inteira do comitê sem qualquer participação, intimidados pelo aparente conhecimento do outro, técnico. Nesse sentido, a escolha dos pais recaiu sobre pais-professores, em uma "contaminação" da categoria pais como *stakeholders*.

Como já mencionado, fizeram parte do escopo da AA, além dos dados buscados por meio de testes aplicados aos alunos de 1ª a 4ª, informações coletadas ao longo da sua implementação, por meio do RD ou ainda de roteiros de entrevistas / questionários aplicados pelos coordenadores de aplicação a diretores ou professores, em amostras controladas. Esses dados foram utilizados para composição dos Relatórios-síntese de monitoramento da AA e para seu refinamento de modo a atender às escolas, segundo texto dos folhetos *Avaliação de Aprendizagem: participação e contribuições das escolas* nos anos de 2002 e 2003. Em algumas ocasiões, foram realizados estudos paralelos como, por exemplo, a investigação sobre o perfil do professor e o desempenho dos alunos, conduzida pela equipe central com a participação de pesquisadores em 2003, cujo resultado pode ser lido no *folder* da Avaliação da Aprendizagem 2003.

Quanto ao item de verificação do U3 sobre novas incorporações, é facilmente observável que a AA foi ampliada para incluir a avaliação da produção textual, no atendimento a uma demanda freqüente das escolas participantes apresentada nas respostas aos RD de 2001 e 2002. As matrizes de produção textual foram criadas pelos mesmos procedimentos das demais e lançadas em 2003, com testes para a 4ª série, expandida em 2004 para incluir a 3ª série, como visto nos capítulos introdutórios.

Dois itens de verificação merecem uma atenção no U3: "obtém informação suficiente para avaliar o mérito do programa" e "obtém informação suficiente para avaliar o valor do programa". Essa atenção é devido ao fato de que, como já discutido, os objetos da avaliação e da meta-avaliação são diferentes. Uma coisa é investigar se a AA teve mérito e valor (meta-avaliação da política); uma coisa diferente é verificar se a AA focalizou mérito e valor do seu objeto, o ensino de 1ª a 4ª nas escolas públicas da Bahia.

De acordo com informações dos relatórios técnicos da AA, em especial aqueles da Psicometria, a coleta de dados nas amostras controladas permitiu a análise do mérito da própria AA. Todos os itens, testes e demais materiais de aplicação tiveram seu comportamento acompanhado[68]. Quanto ao valor da política, as perguntas do RD favoreceram tal análise quando observados *stakeholders* principais (as escolas), ainda que, em 2001 e 2002, esse processo tivesse um caráter exploratório, somente sistematizado a partir de 2003. Dados sobre o valor da AA não foram levantados de maneira sistemática junto aos demais *stakeholders*.

Quando vistos, no entanto, valor e mérito do ensino de 1ª a 4ª, objeto da AA, o cenário encontrado foi diverso. A AA, como ferramenta diagnóstica, não buscou analisar o mérito ou o valor da educação oferecida pela escola (os programas e planos de curso e de aula, por exemplo). A AA focalizou os efeitos da abordagem escolhida (qualquer que tivesse sido) no desempenho dos alunos em Português e Matemática, a cada 200 horas letivas. Como previsto em lei, cada escola deveria ter autonomia pedagógica para escolher a melhor estratégia de ensino. Os testes externos, aplicados após a conclusão de cada unidade, lhe dariam uma medida de quão distante ou perto seus alunos estavam em relação à proficiência esperada. Mesmo os materiais didáticos encaminhados às escolas em resposta à identificação dos descritores para os quais os alunos tinham apresentado os resultados mais insatisfatórios não eram uma resposta específica a uma ou outra abordagem didática. Em reação aos baixos resultados obtidos, a leitura dos relatórios síntese dos RD mostrou que, ao longo do tempo, um número de escolas se manifestou, a cada aplicação, contra o que consideraram inadequação das matrizes para suas realidades, em geral pobres e distantes da capital. Esse movimento foi mais forte no início da AA, mas tais manifestações ocorreram até o final do ciclo, como visto em Dantas (2005).

Diante da situação detalhada acima, uma análise geral do indicador U3 aplicado a AA aponta para:

1) a restrição do uso da avaliação às atividades que envolveram Língua Portuguesa e Matemática, pelo menos no caráter instrumental de uso;
2) não inclusão de análise de mérito das abordagens educacionais nas escolas, traduzindo-as apenas ao desempenho dos alunos nessas duas disciplinas;

68 O presente estudo volta-se para a Utilidade. A categoria Precisão não foi por ele endereçada. Entretanto, os relatórios técnicos referentes a cada aplicação de provas – seja em caráter piloto, seja na amostra controlada – estão disponíveis para outras pesquisas. Para fins da presente investigação, assume-se que a AA foi implementada sob critérios rigorosos que lhe conferem mérito avaliativo.

3) percepção das escolas de não pertencimento pela falta de representatividade nas negociações de escopo ou pela não adequação de matrizes para as realidades distintas, o que também pode indicar comprometimento para o uso e a utilidade;
4) alguma flexibilização da AA para incorporação de novas questões avaliativas;
5) coleta de dados suficiente para o escopo determinado, e
6) esforços condizentes com as prioridades estabelecidas.

Dos dez itens verificadores, oito foram identificados na AA, mas para uma ação restrita à Língua Portuguesa e à Matemática e foco nas escolas. A figura a seguir apresenta a síntese da análise da AA no Indicador U3, categoria Utilidade.

Comportamento da AA quanto aos itens de verificação do indicador U3

- ✓ Entende os requisitos da avaliação mais importantes para o cliente.
- ✓ Entrevista *stakeholders* para determinar suas perspectivas (participação em oficinas).
- ✓ Assegura negociação entre avaliador e cliente sobre públicos pertinentes, questões avaliativas e demanda de informação.
- ✓ Atribui prioridade para os *stakeholders* mais importantes (escolas).
- ✓ Atribui prioridade para as questões mais importantes (dentro do recorte possível de executar).
- ✓ Flexibiliza a adoção de novas questões durante a avaliação (refinamento para a escola apenas).
- ✓ Obtém informação suficiente para atender às questões mais importantes dos *stakeholders*.
  Obtém informação suficiente para avaliar o mérito do programa*.
  Obtém informação suficiente para avaliar o valor do programa*.
- ✓ Aloca esforços avaliativos de acordo com as prioridades determinadas às informações necessárias.

* Para fins da política de avaliação educacional com foco nas escolas, como é o caso da AA, considera-se "programa" o ensino de 1ª a 4ª séries oferecido por cada unidade escolar.

4.1.4 Análise do U4 – Identificação de valores na Avaliação de Aprendizagem.

O indicador U4 trata da base fornecida pela avaliação para valoração dos resultados, em ambiência de respeito às leis, aos objetivos do programa sob análise e aos valores dos *stakeholders*. Quando são considerados os objetivos mais abrangentes da AA, quanto ao ritmo da escola e à redução das taxas de reprovação e aumento das taxas de aprovação, não houve uma definição da linha de corte sobre o patamar aceitável. Especificamente, no entanto, foram fornecidas as bases para o julgamento de valor do desempenho das 1ª a 4ª séries em Português e Matemática. Após a correção dos testes, as escolas fariam sua interpretação dos resultados obtidos e buscariam corrigir os problemas identificados.

Para cada teste, foi definida a linha de corte para o comportamento da turma avaliada, por dimensão da matriz de referência[69] (consideradas as duas disciplinas, as quatro séries e as três unidades letivas para as quais os testes eram encaminhados). De acordo com o Relatório *Determinação de Linhas de Corte para as Escalas de Proficiência de 1ª a 4ª séries da Avaliação de Aprendizagem 2003 – 1ª unidade*, as faixas da escala foram nominadas Proficiente e Não Proficiente e as linhas de corte foram obtidas pelo método Angoff modificado[70], sendo informadas às escolas a cada aplicação. Nesse sentido, a definição de proficiência foi dada por uma metodologia estatística, não havendo, portanto, a consideração do juízo de valor dos *stakeholders (*a ser verificada no U6) quando da sua definição.

Negociações no início da implementação da política de AA reduziram a expectativa original – de identificação do patamar de qualidade para cada descritor da matriz observado no comportamento individual do aluno – para o comportamento aceitável da turma quanto às dimensões de conteúdo nas quais os descritores estivessem alocados. A redução deu-se porque, por um lado, para uma informação individual, o aluno seria submetido a um teste muito longo, não recomendável para séries iniciais. Por outro lado, o apoio à individualização dos resultados poderia conduzir as escolas a substituírem suas avaliações usuais pela avaliação externa, em um efeito colateral danoso para o processo de autonomia que se pretendia no contexto político do Educar para Vencer.

Como já discutido no indicador U3, o escopo foi proposto a partir das orientações legais vigentes, não havendo, portanto, problemas quanto ao item de verificação "leva em consideração as leis pertinentes".

No desenho da AA, os responsáveis pela interpretação valorativa foram os *stakeholders*/usuários (diretores, coordenadores e professores). O problema identificado pela leitura dos relatórios síntese de monitoramento da AA referiu-se não à falta de uma definição do *stakeholder* a julgar o resultado, mas à proposição de proficiência / não proficiência determinada por meio de testes padronizados, de múltipla escolha, considerados por muitos como não contextualizados (especialmente no caso de Matemática). Vários professores e coordenadores pedagógicos (inclusive municipais) reagiram a esse formato e, portanto, ao diagnóstico feito a partir dele.

---

69 Nas matrizes, os descritores foram agrupados em dimensões ou domínios. Por exemplo, para as 1ª e 2ª séries, os domínios avaliados foram pré-leitura, leitura e leitura e escrita, para português; números e operações, espaço e forma, grandezas e medidas e tratamento da informação, para matemática.

70 Para uma análise sobre métodos de determinação de padrão, tais como o Angoff e o Angoff modificado, consultar Ricker, 2003.

Para reforçar o julgamento de valor, a AA atendeu a uma demanda das escolas, apresentada por meio dos RD, e criou os guias-diagnóstico, em 2002. Os guias, dirigidos aos professores, foram encaminhados junto ao material de aplicação. A idéia do Guia - diagnóstico foi contribuir para que os professores refletissem sobre o desempenho de suas turmas, à luz de sua prática em sala de aula e do planejamento de curso e respeitado o limite de escopo sob avaliação (informado nas matrizes de referência), de modo a enriquecer a reunião de coordenação sobre os resultados obtidos e conseqüente replanejamento, quando necessário. O trecho a seguir foi transcrito do *Manual do Professor. Avaliação de Aprendizagem* 2004 – 1ª unidade (p.12).

> **Caro (a) Professor(a),**
> O guia a seguir foi elaborado para ajudá-lo a analisar os resultados de sua turma após o preenchimento dos Quadros-diagnóstico para Português e para Matemática. Pense sobre as questões abaixo antes de participar da reunião para discussão dos resultados na sua escola. Algumas das informações solicitadas devem ser respondidas durante a aplicação da prova; sendo assim, esteja com esse instrumento em mãos nesse momento. Sua contribuição será fundamental para a melhora da qualidade de ensino oferecido na sua unidade.
> Para aproveitar ao máximo este roteiro, tenha em mãos os seguintes materiais:
> - Os Quadros-diagnóstico preenchidos.
> - A Matriz de Referência da Avaliação de Aprendizagem (adequada à série da turma).
> - Seu plano de curso.
> - Alguns exemplares dos testes de Português e Matemática respondidos por seus alunos.
>
> **Caso o espaço para respostas seja insuficiente, continue em outra folha de papel.**

Em síntese, o comportamento da AA observado pelos itens verificadores do indicador U4 apontam para uma base clara e defensável para o julgamento de valor em termos dos conteúdos em Língua Portuguesa e em Matemática, mas em contexto de reação por alguns dos *stakeholders* principais (escolas e, nelas, os professores) ao instrumento utilizado (os testes padronizados). Isso é importante porque foram esses *stakeholders*/usuários que ficaram responsáveis pelas interpretações dos achados e pelas recomendações posteriores. Os objetivos gerais da avaliação – ritmo para as escolas e queda da reprovação -, no entanto, não foram acompanhados e tampouco o foram as abordagens de ensino-aprendizagem adotadas pelas escolas. As necessidades sociais e dos *stakeholders* foram consideradas na medida em que estiveram envolvidas com o ensino de Língua Portuguesa e Matemática. Dos dez itens verificadores, considera-se que a AA respondeu positivamente a quatro e que dois não sejam aplicáveis. Mais uma vez, essa análise esteve restrita às escolas.

Comportamento da AA quanto aos itens de verificação para o indicador U4

Considera fontes alternativas de valores para interpretação dos achados da avaliação. (Não se aplica; a própria escola faz sua interpretação à luz da sua realidade).
✓ Fornece uma base clara e defensável para os julgamentos de valor.
✓ Determina a parte apropriada para fazer as interpretações valorativas.
Identifica necessidades sociais pertinentes (AA parte de uma matriz comum).
Identifica necessidades pertinentes dos usuários (AA limita-as a Português e Matemática).
✓ Leva em consideração as leis pertinentes.
✓ Leva em consideração a missão institucional (no caso, os objetivos do Governo/SEC).
Leva em consideração os objetivos do programa (leva em consideração os conteúdos em Port e Mat e não os conteúdos trabalhados pela escola).
Leva em consideração os valores dos *stakeholders* (leva em consideração os resultados do Angoff).
Apresenta interpretações alternativas fundamentadas em base valorativa crível, ainda que conflitante. (Não se aplica; a própria escola faz sua interpretação à luz da sua realidade).

4.1.5 Análise do U5 – Clareza no relato da Avaliação de Aprendizagem

No presente estudo, o indicador U5 compreende itens de verificação da comunicabilidade dos relatos da avaliação, especialmente em termos de conteúdo e linguagem. Considera-se, ampliando os usos da avaliação para além dos resultados, a análise das comunicações entre os avaliadores e os *stakeholders* e não somente dos relatórios de resultados[71]. Na política de avaliação em tela, foram as próprias escolas a aplicar os testes, corrigi-los e, a partir das respostas dos alunos, compor os quadros-diagnóstico. As comunicações como matrizes, manuais e guias foram, portanto, elementos de facilitação de uso de processo tanto quanto de resultado.

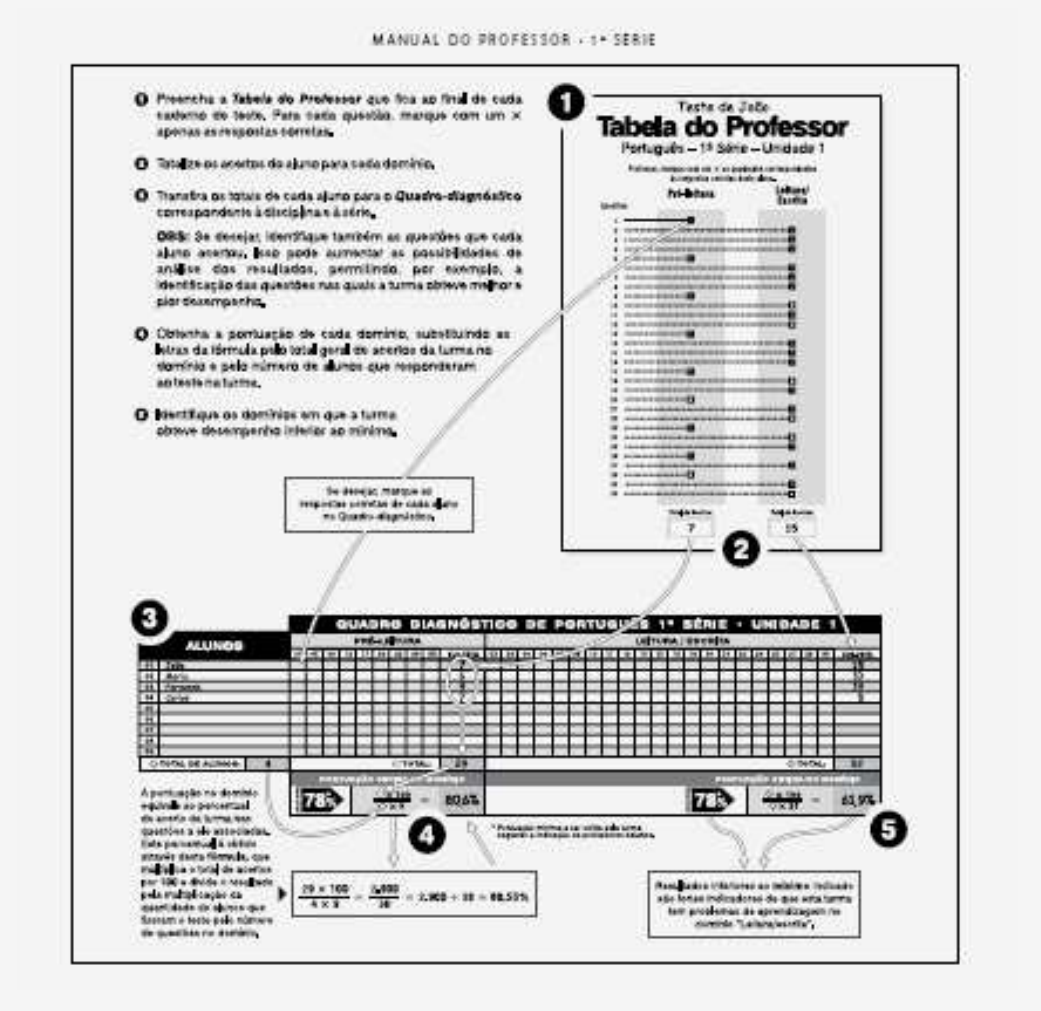
71 A relação de documentos analisados encontra-se no capítulo Metodologia (Subseção 3.3).

Ilustração 15: Exemplos de materiais encaminhados às escolas pela equipe central da AA em 2004 (capa da matriz de referência 1ª e 2ª séries, capa da matriz de produção textual 4ª série, capa do manual de pré-teste de um teste de produção textual para a 4ª série, capas dos vídeos de remediação).

No caso da AA, a análise dos materiais encaminhados às escolas (das matrizes até os manuais e materiais de remediação) permite observar uma identidade visual que os distingue, desde o início em 2001, como pode ser visto nos exemplos apresentados na ilustração acima. Assim como ocorreu mais tarde com a Prova Brasil, mostrada na Subseção 2.4.3.5, foi feita uma escolha por materiais e figuras estilizadas e leves, coloridas (quando possível, já que a impressão das capas das provas em cores, por exemplo, encareceria muito mais o custo de cada aplicação) e que pudessem facilmente ser transformadas em cartaz para utilização pelas escolas e pelas secretarias. Até o momento, somente a AA e a Prova e Provinha Brasil adotaram essa estratégia de comunicação. Mesmo um estado como o Paraná, que emitia um Boletim da Escola ao final das suas avaliações, preferiu o formato padrão dos relatórios clássicos de avaliação.

Em termos da linguagem utilizada, foi adotado um padrão direto, com parágrafos curtos e linguagem objetiva em todas as peças de comunicação avaliação–escola e avaliação–secretarias

municipais de educação. A análise dos manuais de aplicação e correção permitiu observar que os mesmos traziam orientações com exemplos ilustrativos, como demonstrado na figura a seguir[72].



Fonte: Projeto de Avaliação Externa. Manual do Professor para o teste operacional de Língua Portuguesa – 1ª série – 1ª unidade 2004

Ilustração 16: Exemplo de utilização de imagens como reforço à comunicação com as escolas. Em tela, uma orientação para preenchimento do Quadro Diagnóstico.

Quando houve necessidade de comunicação de aspectos técnicos, como o conceito de descritor, nas matrizes de referência, ou de operadores de coerência, na matriz de produção textual, os textos trouxeram um esclarecimento sobre os mesmos. As matrizes de referência apresentaram exemplos para que o leitor-escola pudesse identificar o descritor, como mostra a figura a seguir, importada da matriz de produção textual.

72 No caso da 1ª série, os manuais ofereciam um roteiro de aplicação, a ser lido pelo professor responsável pela turma de modo a padronizar a aplicação dos testes e a contribuir para a qualidade da informação coletada.

Descritor:	**Produzir um pequeno texto narrativo, baseando-se na ilustração e observando a clareza, a coesão e os marcadores textuais.**
Questão exemplo:	Escreva um pequeno texto narrativo, tendo como base a gravura.
Resposta:	Final de Campeonato Carlinhas era o melhor jogador do time de futebol da nossa escola. Mas ele quebrou a perna numa queda e ficou vários meses sem jogar. O time da escola começou a perder várias partidas. No dia da final do campeonato Carlinhas decidiu jogar. Todos ficaram com medo. Será que ele tinha se recuperado? Mas foi uma festa. Carlinhas fez 5 gols e o nosso time levou a tassa de campeão! 
Considerações:	O aluno precisou usar sua habilidade da observação e análise, para elaborar uma narrativa criativa relacionada à gravura. Como mostra o exemplo, é natural que nesta idade os alunos mostrem dúvidas quanto ao emprego de "ç" e "ss" ("tassa" no lugar de "taça"), que devem ser corrigidas.

Fonte: Projeto de Avaliação Externa. Matriz de Referência de Produção Textual – 4ª série – 2003 (p. 60)

Ilustração 17: Descritor exemplificado na Matriz de Produção Textual 4ª série

O uso da AA esteve atrelado ao entendimento que a escola pudesse ter tido sobre como aplicar e como corrigir os testes e, a partir daí, reajustar seu plano de curso e os planos de aula dos seus professores para endereçar os problemas diagnosticados. Nesse sentido, desde 2001 os RD incluíram questões sobre a adequação da linguagem. Em 2002, o *RD* da 3ª unidade incluiu algumas perguntas para levantar a percepção das escolas sobre a linguagem utilizada nas publicações e manuais. De acordo com dados do Relatório Síntese de Monitoramento da AA naquela unidade, do total de 2.756 escolas envolvidas, 1.414 escolas (429 estaduais e 985 municipais) encaminharam o RD à equipe central, respondendo a essas perguntas. Dentre elas, 716 escolas consideraram a linguagem "clara" e 328 unidades marcaram "clara" e "detalhada" simultaneamente. Uma apreciação negativa foi identificada em apenas 32 relatórios, para os quais a linguagem foi confusa ou redundante.

Dados sobre a linguagem foram novamente coletados em 2004, nos três RD, e estão apresentados na tabela a seguir. Nesse ano, um número muito grande de escolas simplesmente não relatou sobre a clareza e o nível de linguagem dos materiais da AA (exceto as provas), mas apenas três a consideraram inadequada, dentre os 182 relatos válidos.

Tabela 3: Posição das escolas quanto à linguagem utilizada pela AA em suas comunicações (exceto provas) em 2004.

Classificação da posição da escola	Freqüência	Percentual Válido
Não relata	1.166	86,5
Linguagem inadequada	3	0,2
Linguagem adequada	179	13,3
Total	1.348	100,0

Fonte: Base de dados RD 2004.

Ainda sobre aspectos de comunicabilidade com o *stakeholder* / usuário principal, uma demanda encaminhada em RD para a Avaliação referia-se ao tamanho das letras e ao espaço de preenchimento dos quadros-diagnóstico[73], antes desenhados para uma folha A4. Para atendimento da demanda da escola, os quadros passaram a ser impressos com dobra, de modo a aumentar o espaço de preenchimento e facilitar a leitura.

As questões sobre linguagem dos relatos e comunicações, discutidas até aqui, referiram-se aos *stakeholders*/usuários das escolas, foco do presente estudo. Os demais *stakeholders* receberam materiais de divulgação, especialmente com os calendários de aplicação das provas, e, em 2004, os relatórios-síntese, com as informações coletadas através dos RD e com o panorama de desempenho dos alunos da amostra monitorada. Desde 2002, DIRECs e secretarias foram incluídas no processo de distribuição de materiais de aplicação, de modo que pudessem acompanhar as aplicações realizadas em seus municípios ou regiões. Mantendo as características de comunicação acima relacionadas, o Projeto de Avaliação Externa disponibilizou um *site* em 2004 (www.aval.ufba.br). Não há relatos sobre a utilização dessa ferramenta nos documentos analisados e a última atualização feita data de agosto do mesmo ano.

O cliente SEC/SPDE-SUPAV[74] recebeu, desde a primeira aplicação, além de cópia de todos os materiais encaminhados às escolas, relatórios técnicos sobre a construção dos instrumentos e das escalas e sobre o monitoramento nas escolas (amostra) e o desempenho dos alunos nas duas disciplinas avaliadas. Recebeu também os relatórios de logística, nos quais eram relacionados os problemas e as soluções postas pelas escolas quanto a atraso, falta ou sobra de materiais, comportamento da empresa transportadora, adequação de aplicação de instrumentos, dentre outras

73 Tabelas inseridas no Manual do professor, por série e por disciplina, para qual deveriam ser transpostas as respostas de cada aluno de modo que o panorama de desempenho da turma, por dimensão avaliada, pudesse ser obtido.

questões. O quadro a seguir relaciona tais relatórios, além de alguns relatórios anuais. Esses relatos mais técnicos, ainda que tivessem mantido um padrão de comunicação direta, apresentaram linguagem diferente daquela utilizada no cotidiano da sala de aula.

Relatório Técnico	Setor Responsável	Freqüência	*Stakeholder*
Análise psicométrica das provas de Português	Psicometria	Unidade avaliada	Cliente / Academia
Análise psicométrica das provas de Matemática	Psicometria	Unidade avaliada	Cliente / Academia
Determinação de linhas de corte para as escalas de proficiência de 1ª a 4ª séries da Avaliação de Aprendizagem	Coordenação de Desenvolvimento da Aprendizagem	Unidade avaliada	Cliente
Síntese da Logística	Coordenação de Administração de Instrumentos	Unidade avaliada	Cliente / Interveniente
Síntese da categorização e organização dos dados do Relatório do Diretor (até 2003)	Coordenação Geral / em 2004, Coordenação de Desenvolvimento da Aprendizagem	Unidade avaliada	Cliente / Secretários municipais
Síntese da Unidade (em 2004, substituindo o anterior)	Coordenação de Desenvolvimento da Aprendizagem	Unidade avaliada	Secretarias municipais e DIREC
Síntese da categorização e organização dos relatórios do Coordenador de Aplicação	Coordenação de Administração de Instrumentos	Unidade avaliada	Cliente / Interveniente no Convênio
Vídeos Didáticos e Material de Apoio	Coordenação de Desenvolvimento da Aprendizagem	Unidade avaliada	Cliente
Gerencial de comportamento de itens	Psicometria	Anual	Cliente
Panorama Geral - Vídeos Didáticos e Material de Apoio	Coordenação de Desenvolvimento da Aprendizagem	Anual	Cliente

Quadro 13: Panorama dos tipos de relatórios utilizados pela Avaliação de Aprendizagem na comunicação dos seus aspectos técnicos durante sua implementação no ciclo 2001-2004.

Na análise de aspectos da comunicabilidade que possam impactar de alguma maneira o uso da avaliação, há ainda uma questão que merece reflexão: o grau de detalhamento dos relatórios técnicos e mesmo dos relatórios síntese e o conseqüente tamanho do texto. É senso comum que a maior parte dos dirigentes não dedica muito tempo à leitura de relatórios, o que requer um esforço grande de síntese. Há, entretanto, aspectos essenciais que precisam constar de relatórios, inclusive aqueles que remetem aos limites das interpretações dos resultados, como os recortes de escopo, o intervalo de confiança dos dados, as possibilidades de comparabilidade, dentre tantos (ver crítica ao boletim da escola na Prova Brasil, apresentado na seção 2.4.2.5). Os relatórios técnicos da AA

74 Aos demais stakeholders não foram encaminhados relatórios técnicos da AA.

se valeram de tabelas e gráficos, além de apresentarem glossários. Nesses relatórios não foram endereçadas questões contratuais.

Diante desse panorama, a verificação dos itens no Indicador U5 na AA aponta para favorecimento do uso da avaliação. A linguagem foi simples e direta e os materiais foram encaminhados às escolas, que se encarregaram de produzir seus próprios resultados (não havendo, portanto, relatórios de resultados a encaminhar para elas). Os demais *stakeholders* foram contemplados com relatos síntese, mais técnicos que os materiais enviados às escolas, mas com igual cuidado no formato e na linguagem. Não foi observada utilização de sumários executivos. O material da AA analisado incluiu uma série de apresentações em *power point* preparadas para públicos os mais diversos, inclusive para a SEC e para a academia que, de algum modo, sumarizavam os resultados da AA por aplicação do estudo monitorado/ano. Dos dez itens verificadores, a AA correspondeu positivamente em oito deles, de maneira ampla, como pode ser visto na síntese a seguir.

Comportamento da AA quanto aos itens de verificação para o indicador U5

- ✓ Relata de maneira clara as informações essenciais.
- ✓ Divulga relatórios breves, simples e diretos.
  Focaliza relato das questões contratuais.
- ✓ Descreve o programa e seu contexto.
- ✓ Descreve os propósitos da avaliação, seus procedimentos e achados.
- ✓ Fundamenta conclusões e recomendações (por meio de quadros e guias-diagnóstico).
- ✓ Evita utilização de jargão técnico (especialmente nas comunicações com as escolas).
- ✓ Utiliza a linguagem dos *stakeholders* nos relatos (na comunicação com as escolas).
  Fornece sumário executivo.
- ✓ Fornece relatório técnico (à SEC).

4.1.6 Análise do indicador U6 – Tempo e divulgação dos relatórios da Avaliação da Aprendizagem

O U6 trata da perecibilidade dos resultados da avaliação e dos esforços da equipe avaliadora no sentido de fazer a informação chegar, no tempo certo, ao cliente, aos demais *stakeholders* e usuários. A definição sobre a conveniência do tempo está nas mãos daqueles que vão usar a avaliação.

No caso da política sob análise – a AA – esse aspecto da conveniência do tempo foi crucial para a utilização dos dados, visto que era o *stakeholder*/usuário principal – a escola – não só quem deveria tomar decisões embasadas nos dados, mas quem iria produzi-los a partir de procedimentos e orientações encaminhados a cada unidade letiva. Ficou patente, nessa experiência, a defesa de Ferrer (1997) sobre a necessidade de fluxo contínuo de informações nas avaliações. A figura a seguir traz um cartaz utilizado pela AA em 2004 para apresentar essa noção de fluxo da avaliação, ao tempo que chamava atenção para os tempos recomendados para o uso das informações especialmente atrelado ao replanejamento no início de cada nova unidade letiva.

Ilustração 18: Cartaz utilizado pela AA, em 2004, para comunicar a noção de continuidade do fluxo de informações da avaliação na escola, *stakeholder*/usuário principal.

Como já mencionado, pelo delineamento da AA (ver Relatório de Conclusão do Convênio 444/99), no início de cada ano, as matrizes de referência eram divulgadas e recomendava-se que as comunidades escolares reunissem-se para analisá-las no contexto de seus planos de curso. A cada unidade já concluída (até a 3ª), os professores das séries avaliadas deveriam aplicar os testes enviados pela AA. A correção desses instrumentos, feita pelos próprios professores, deveria ser finalizada com o preenchimento de um quadro diagnóstico por turma, de onde constavam as informações sobre os cálculos para obtenção dos resultados e os percentuais mínimos de proficiência esperados. Em seguida, esses professores deveriam responder a um guia-diagnóstico (já mencionado, encaminhado também no Manual do Professor). Cabia ao coordenador pedagógico (ou, na sua ausência, ao diretor da escola) convocar uma reunião dos professores para discutir os achados e propor ações para correção de rumos, quando necessário. A síntese dessa reunião (acrescida de comentários, sugestões e críticas à AA) deveria fundamentar as respostas

para o RD, encaminhado em formulário próprio de volta à equipe central da avaliação. Pelas informações do relatório síntese de monitoramento da 3ª unidade de 2004, dentre as escolas que enviaram o RD a tempo, aproximadamente 90% relataram ter tido a reunião, o que aponta para uma alta possibilidade de uso da AA para esse grupo.

No referido delineamento, como a própria escola produzia resultados imediatos, não havia necessidade de encaminhamento de relatórios parciais ou final[75]. Como foi discutido no U5, o esforço de comunicação foi concretizado a partir do envio dos materiais às escolas, no sentido das orientações para a aplicação, para correção e para análise dos dados obtidos. Os problemas na logística de distribuição dos materiais, em termos do tempo, estiveram relacionados às diferenças nos calendários letivos das escolas das diversas redes, que dificultavam a entrega dos materiais das AA ao final de cada unidade letiva. Os calendários das redes municipais nem sempre acompanharam o calendário da rede estadual e uma série de características locais, como colheitas, feiras, festas, atreladas ou não ao período de férias no meio do ano, interferiu em sua implementação. Em 2004, de acordo com informações dos Relatórios Síntese, houve algum problema de atraso da entrega dos materiais nas escolas (devido a processos licitatórios longos) e houve atraso na aplicação das provas devido às greves das redes estadual e municipais. Tais atrasos não contribuíram para o uso da AA (ou mesmo para a aplicação da AD, como discutido no capítulo 3. Metodologia).

Quanto aos demais *stakeholders*, havia o Relatório Síntese da AA em cada unidade de aplicação. Até 2003, esse relatório era encaminhado apenas ao cliente. Em 2004, em resposta a uma demanda das escolas, passaram também a ser encaminhados às secretarias municipais e às DIREC. Como, para elaboração dos Relatórios Síntese, era necessário que as escolas tivessem encaminhado seus RD preenchidos para a equipe central, essa tarefa não era cumprida de maneira imediata. Por essa razão, os relatórios síntese não contribuíram diretamente para que os órgãos centrais ou mesmo suas representações regionais apoiassem as escolas em suas ações de remediação ao longo da unidade letiva seguinte ao diagnóstico. A equipe central de avaliação (o *staff*), por outro lado, era constantemente alimentada com informações encaminhadas por correspondências formais dos órgãos centrais da educação estadual e municipais e, especialmente, pelas informações contidas nas respostas ao RD desde 2001.

75 No Brasil, à época, apenas o Paraná adotou a estratégia de correção pelas escolas no caso de questões de elaboração de texto.

Na análise da AA quanto aos itens de verificação da interação equipe - *stakeholder*, percebeu-se um esforço de fechamento das atividades/conclusão ano a ano. Desde 2002, a equipe de avaliação passou a encaminhar às escolas e secretarias um folheto intitulado *Avaliação de Aprendizagem – Participação e contribuições das escolas*. Essa publicação foi criada para promover um fechamento anual das ações da AA e nela eram divulgadas as percepções das escolas, esclarecimentos para as dúvidas mais freqüentes apresentadas ao longo do ano, e relato de algumas das propostas de trabalho implementadas pelas escolas no combate aos problemas diagnosticados. O folheto foi encaminhado no final do ano ou no início do ano seguinte e também informava escolas e secretarias sobre o calendário de provas do ano seguinte. Comunicações individuais com escolas e secretarias foram feitas através de ofícios da equipe de AA. Para a sensibilização dos *stakeholders* quanto à avaliação, a equipe de AA participou dos fóruns coordenados pelas equipes de Capacitação Gerencial das Unidades Municipais e de Fortalecimento da Gestão Escolar, por todo o Estado da Bahia, especialmente nos anos de 2001 a 2003.

Não há registro de interação com a imprensa em termos da AA, exceto em Diário Oficial do Estado para divulgação do calendário letivo. Não há registro sobre utilização da TV. A *internet* foi utilizada a partir de 2004, para informações gerais sobre o Projeto de Avaliação Externa (com *links* para o ISP/UFBA e para o Educar para Vencer, na página da SEC)[76].

A análise da AA quanto aos itens verificadores do U6 aponta para uma predominância de informações para o *stakeholder*/usuário principal, em tempos convenientes, mas para um não atendimento dos demais *stakeholders* a tempo de apoiarem as escolas durante o ano letivo. A AA não contou com divulgação pela imprensa e a comunicação foi feita diretamente com os *stakeholders*, principalmente por meio dos materiais de aplicação e correção dos testes. Dos dez itens, a AA respondeu positivamente a cinco deles, como registrado no quadro a seguir.

76 Ver www.aval.ufba.br, ainda disponível em junho de 2009.

Comportamento da AA quanto aos itens de verificação do U6

Apresenta relatórios parciais aos usuários-alvo (não se aplica).
Entrega o relatório final quando é necessário (em termos do Relatório Síntese, somente em 2004)
✓ Interage em tempo conveniente com os *stakeholders* / usuários principais.
✓ Interage em tempo conveniente com o *staff* da AA.
Interage em tempo conveniente com os demais *stakeholders* (os *stakeholders* fora da escola receberam relatórios muito tempo depois das unidades letivas).
Interage em tempo conveniente com os públicos interessados.
Interage em tempo conveniente com os meios de comunicação.
✓ Emprega mídia adequada para alcançar e informar os diferentes públicos.
✓ Mantém breves as apresentações.
✓ Utiliza exemplos para ajudar os públicos a relacionar os achados com situações práticas.

4.1.7 Análise do U7 – Impacto da Avaliação de Aprendizagem.

O último dos indicadores na categoria Utilidade oriundos do *checklist* de Stufflebeam para a proposta do JCSEE lida com as estratégias utilizadas para o estabelecimento da relação avaliador x *stakeholders*, no sentido de potencializar os usos da avaliação. Como mencionado na seção 2.4.2.7, também esse indicador é verificado por uma série de itens de alguma maneira tratados ou tangenciados nos indicadores anteriores. De qualquer maneira, a análise da AA aponta para um uso baixo em relação aos itens de verificação do U7, como será argumentado a seguir.

O primeiro item de verificação trata do contato que a avaliação mantém com seu público-alvo ou seu *stakeholder* principal. No caso da AA, esse contato foi estabelecido principalmente por meio dos materiais encaminhados às escolas, secretarias e DIREC (elaborados com foco na escola). Foram infreqüentes os contatos diretos AA – escola no sentido de visitas da equipe avaliadora, exceto no caso das amostras monitoradas. Houve algumas ações isoladas de capacitação dos docentes em uma determinada unidade (como a Escola Nogueira Passos, rede estadual de Salvador, em 2002 e em 2003) ou de redes inteiras (como Paulo Afonso em 2002 e São Francisco do Conde em 2003). Em termos do órgão central, a AA foi discutida nos vários encontros promovidos pela SPDE, SUPEN ou projetos parceiros (e depois pela SUPAV, Superintendência de Desenvolvimento da Educação Básica - SUDEB e COPE). A equipe central manteve linhas telefônicas disponíveis para o recebimento de ligações (a cobrar) das unidades escolares, para esclarecimento de dúvidas de aplicação e correção (encaminhadas para a Coordenação de Administração de Instrumentos) ou quanto ao conteúdo das matrizes e dos testes (repassadas para os Núcleos de Língua Portuguesa ou de Matemática).

Dúvidas encaminhadas por meio dos RD foram respondidas após categorização desses documentos, por meio de ofícios. Ao longo da AA, o RD foi o canal aberto para o relacionamento com as escolas. Havia sempre perguntas voltadas para o entendimento da escola quanto às orientações fornecidas ou o preenchimento dos formulários encaminhados. Houve um refinamento dos materiais de comunicação AA – escola decorrente desse *feedback*.

Da leitura de qualquer um dos materiais utilizados pela AA, especialmente o Manual do Diretor e o Manual do Professor, fica claro o incentivo ao uso dos resultados no replanejamento escolar e a busca de um ritmo de exposição dos alunos aos conteúdos programáticos (especialmente aqueles constantes nas matrizes de referência, mas sempre com ressalvas de que aquela escolha era limitada por questões de técnica, tempo e custo). Os guias-diagnóstico deveriam cumprir esse papel junto aos professores, enquanto o RD o faria com os coordenadores pedagógicos e/ou com os diretores. Os folhetos encaminhados às escolas e secretarias ao final de 2002 e 2003 (Participações e contribuições das escolas) divulgavam ações (como formação de banco de textos, por exemplo) que as escolas relatavam implementar como estratégia para solucionar os problemas identificados.

Em relação ao apoio para a concretização dos planos de cursos e aulas das escolas, o cenário foi outro. Os vídeos didáticos encaminhados às escolas estaduais e às secretarias municipais de educação deveriam ter servido como reforço às estratégias de remediação propostas pelas escolas. Esse material foi desenvolvido por uma equipe de pedagogos em São Paulo, a partir dos descritores mais críticos em termos de desempenho dos alunos. Os pedagogos escreveram roteiros de aulas-filmes que foram realizados na Bahia, com professores e alunos da rede pública baiana. Para potencializar a utilização dos vídeos, a equipe paulista elaborou um manual para cada série de vídeos, que os acompanhava. Essa ação, no entanto, teve um impacto restrito por duas razões: as escolas nem sempre tinham os equipamentos de TV e vídeo e as secretarias municipais muito freqüentemente não disponibilizaram os vídeos para suas redes.

A leitura dos Relatórios Síntese permitiu observar que as escolas solicitavam apoio freqüente da equipe de avaliação em relação à dificuldade em lidar com os conteúdos constantes nas matrizes (especialmente produção textual e subtração na 3ª série); em coordenar os professores; ou em definir ações a partir do diagnóstico feito. À exceção dos vídeos didáticos, não houve, no delineamento da AA, uma ação voltada para esse apoio específico, especialmente porque essa política não foi formulada como uma intervenção além da avaliação. Como discutido

anteriormente, há uma tendência à demanda da equipe avaliadora como equipe gestora ou ainda como interventora, o que não é recomendável.

Os RD trouxeram grande demanda para capacitação em aspectos de Língua Portuguesa e Matemática. Houve ainda demanda para capacitação em gestão de turmas heterogêneas (ou "difíceis"), avaliação, assuntos gerais, dentre outros. Essas solicitações foram encaminhadas ao cliente (em reuniões específicas) e às secretarias municipais de educação por meio dos Relatórios Síntese (apenas em 2004). As secretarias (incluindo a SEC), de maneira geral, não contavam com recursos suficientes para que suas equipes pedagógicas garantissem esse apoio.

Dos dez itens de verificação do indicador U7, foi possível observar um comportamento positivo da AA em quatro, sendo os demais parcialmente atendidos ou não atendidos. Houve um esforço de envolvimento com os *stakeholders* no desenvolvimento da implementação da AA, mas não havia previsão, no delineamento da ação, de suporte aos *stakeholders* para a mudança, de maneira isolada. A AA não se mostrou articulada o suficiente, em termos das secretarias e da SEC, para que o apoio fosse oferecido pelos órgãos competentes. A possibilidade de utilização da AA, nesse cenário, pode ser considerada baixa.

Comportamento da AA quanto aos itens de verificação do U7

- ✓ Mantém contato com o público-alvo (escola como *stakeholder* principal).
  Envolve *stakeholders* ao longo da avaliação (concentra-se nas escolas).
  Incentiva e apóia *stakeholders* na utilização dos achados (parcialmente).
  Demonstra aos *stakeholders* como utilizar os achados em sua prática/trabalho (parcialmente).
- ✓ Prevê e endereça usos potenciais dos achados.
  Provê relatos parciais (não se aplica).
- ✓ Assegura que os relatórios sejam abertos, francos e concretos.
- ✓ Suplementa comunicação escrita com comunicação oral contínua.
  Conduz *workshops* de *feedback* para rever e aplicar os achados.
  Organiza-se de modo a oferecer *follow-up* aos usuários na interpretação e utilização dos achados.

Sendo o U7 o último indicador na categoria Utilidade, de acordo com a proposta de adaptação dos padrões do JCSEE e do *checklist* de Stufflebeam (1999), a apontar as possibilidades de uso da avaliação, a subseção a seguir sumariza as discussões anteriores em um quadro geral da categoria para, em seguida, serem apresentados os resultados para o 8º indicador, percepção de utilidade.

4.1.8 Síntese do comportamento da Avaliação de Aprendizagem nos indicadores da categoria Utilidade adaptada do JCSEE

A categoria Utilidade, com seus sete indicadores, foi proposta pelo JCSEE para meta-avaliação dos elementos da avaliação de programas educacionais que assinalassem sua qualidade em relação à utilização. Esses indicadores foram avaliados por itens de verificação propostos por Stufflebeam (1999) em um *checklist* de meta-avaliação. As sete subseções anteriores apresentaram o comportamento da AA a partir da análise sobre a presença x ausência desses elementos, adaptados para a análise de políticas de avaliação (em lugar de programas de educação). No *checklist* original, cada indicador seria analisado por meio de dez itens verificadores. Após a adaptação para análise de políticas de avaliação, o número de itens verificadores variou, como pode ser visto no quadro a seguir.

Indicadores na categoria Utilidade	Nº de itens do *checklist* original	Nº final de itens verificadores
U1	10	10
U2	10	10
U3	10	10
U4	10	8
U5	10	10
U6	10	9
U7	10	9

Quadro 14: Número inicial e final de itens verificadores nos indicadores da categoria Utilidade

Considerada a escala proposta na Metodologia (Passo III), o quadro a seguir sintetiza os resultados da AA nos sete primeiros indicadores da categoria Utilidade.

Indicador na categoria Utilidade	Nº de itens atendidos	%	*Stakeholder*(s) mais bem atendido(s)	Possibilidade de uso a partir da análise do indicador para a AA
U1	5 (entre 10)	50	Cliente / escola	Média
U2	9 (entre 10)	90	Cliente / academia	Alta
U3	8 (entre 10)	80	Escola	Alta
U4	4 (entre 8)	50	Escola	Média
U5	9 (entre 10)	90	Escola	Alta
U6	5 (entre 9)	55	Escola	Média
U7	4 (entre 9)	44	Escola	Média

Quadro 15: Síntese da análise da AA 2001-2004 pelos itens verificadores da categoria Utilidade, adaptados à análise de política pública para fins do presente estudo.

Nesse panorama, poder-se-ia prever uma utilização da AA de média a alta pelos *stakeholders* principais (escolas e cliente). Essa expectativa contraria os relatos sobre baixa utilização na literatura e deve ser, portanto, confrontada com os usos relatados.

A categoria Utilidade e os seus primeiros sete itens foram entendidos, de alguma maneira, como elementos de qualidade da avaliação que, quando presentes, apontariam para seu uso pelos diversos *stakeholders*. Tais indicadores e itens poderiam ser aplicados em meta-avaliação diagnóstica, ainda quando da formulação da avaliação primária, ou em meta-avaliação somativa. No presente estudo, à categoria Utilidade foi acrescido um oitavo indicador: a percepção da utilidade que os *stakeholders* constroem sobre a experiência avaliativa. Antes de discutir os usos no contexto da AA, a próxima subseção apresenta o registro sobre a percepção que as escolas manifestaram sobre sua utilidade, em si um facilitador do uso.

4.1.9 U8: A percepção de utilidade da AA relatada pelas escolas

Para análise da percepção de utilidade relatada pelas escolas, o presente estudo recorreu às informações coletadas por meio dos RD encaminhados às escolas na 3ª unidade de 2003 e na 3ª unidade de 2004. Nessas duas ocasiões, foram feitas perguntas diretas sobre a percepção da escola sobre os efeitos da AA em algumas áreas, o que permitiu levantar a utilidade da política.

A. Percepção em 2003

Na terceira unidade de 2003, o RD incluiu uma grade com as 20 áreas nas quais as escolas haviam informado, de maneira espontânea nos RD anteriores, alguma percepção de efeito da AA. Os diretores foram convidados a registrar sua percepção em uma escala de cinco níveis: muito positivo, positivo, sem efeito, negativo e muito negativo. Contavam ainda com a opção “não se aplica”.

Das 290 escolas que preencheram a grade, a grande maioria respondeu positivamente em 18 das 20 áreas, com pouquíssimas respostas negativas ou sem efeito. Duas áreas, entretanto, tiveram tantas respostas positivas quanto negativas e sem efeito. A tabela a seguir apresenta a síntese do percentual das respostas positivas, distribuídas por série, quando observado o total de 290 escolas. O RD 2003-3ª unidade era finalizado com uma questão aberta na qual o Diretor deveria justificar sua escolha para as três áreas mais impactadas pela AA. Os registros mais freqüentes, presentes no Relatório Síntese, ilustram a discussão dos dados apresentados.

Tabela 4: Tabela síntese dos percentuais de respostas positivas de 290 escolas, distribuída nas 4 séries do Fundamental Menor, sobre possíveis efeitos da AA para 18 das áreas no RD de 2003 – 3ª unidade.

	Áreas	Efeito da Avaliação de Aprendizagem			
		1ª série	2ª série	3ª série	4ª série
1	Qualidade geral do ensino oferecido aos alunos.	99,3	99,0	63,1	100,0
2	Planejamento de aulas realizado pelos professores.	99,7	98,6	62,1	100,0
3	Avaliações normais da própria escola	95,9	96,9	59,0	98,3
4	Práticas pedagógicas dos professores.	95,9	96,2	59,3	98,3
5	Identificação de problemas de aprendizagem apresentados pelos alunos	95,2	95,5	58,6	96,6
6	Correção de problemas de aprendizagem apresentados pelos alunos	95,5	94,8	57,6	96,9
7	Planejamento de curso realizado pela escola.	94,1	94,8	58,3	96,2
8	Relação dos professores com avaliações em geral.	94,8	94,5	59,7	94,5
9	Desempenho dos alunos em Português durante 2003	94,1	94,8	56,9	97,2
10	Relação dos alunos com avaliações em geral.	93,8	94,1	60,0	95,2
11	Desempenho dos alunos em Matemática durante 2003	91,7	94,5	55,5	95,2
12	Comunicação entre a direção e os professores.	92,1	92,1	60,7	92,4
13	Comunicação entre a escola e seus alunos.	91,4	91,7	61,4	91,7
14	Aprovação dos alunos para as séries seguintes.	88,6	90,7	56,6	93,4
15	Nível de leitura e compreensão de texto apresentados pelos alunos	89,3	89,7	56,2	92,4
16	Desempenho dos alunos em disciplinas não avaliadas durante o curso	84,8	87,6	49,0	87,6
17	Comunicação entre a escola e os pais.	81,7	81,7	51,7	81,0
18	Comunicação entre coordenadores pedagógicos e professores.	71,0	67,6	41,0	67,9
19	Interesse dos pais sobre o desempenho escolar dos alunos.	54,8	55,5	31,0	57,6
20	Participação dos pais nas atividades escolares.	51,4	52,8	26,6	53,8

Fonte: Relatório Síntese Avaliação de Aprendizagem 2003 – 3ª unidade (p. 20).

As maiores freqüências nas respostas atrelaram o efeito da AA à qualidade geral do ensino, ao planejamento de aulas (mais que para o planejamento de cursos), à avaliação da própria escola, às práticas pedagógicas dos professores, e à identificação de problemas de aprendizagem e sua correção (ainda que não necessariamente tivessem contribuído para as melhorias de desempenho dos alunos).

As duas áreas com a maior freqüência de respostas positivas (qualidade do ensino e plano de aula) correspondem ao discurso da AA. Esse quadro pode refletir uma resposta ritualizada, como discutido anteriormente na subseção sobre Usos, especialmente pelo RD ser considerado o instrumento formal de comunicação da escola com a equipe central de avaliação. Chama atenção, por exemplo, que embora um grande número de respostas considere que a AA teve efeito positivo sobre a qualidade geral do ensino oferecido pelas escolas para a 4ª série (100%), esse percentual cai quando são observados o desempenho dos alunos em matemática (95,2%) ou, mais especificamente, em leitura (92,4%).

Quanto à leitura, algumas escolas a mencionaram como área de maior impacto com as seguintes justificativas: “as crianças têm acesso a palavras novas e entendem melhor o que lêem com os materiais da AA” (6 escolas); a AA “possibilita ver claramente a leitura e escrita dos alunos” (4), “porque foi possível detectar que alunos ainda tem dificuldade de leitura e compreensão de texto (1), e “Através dos textos sugeridos e as interpretações, a escola tem descoberto novas maneiras de se trabalhar textos” (1). Interessantemente, o efeito da AA sobre o ensino de outras disciplinas foi justificado por uma das escolas exatamente pelo uso dos textos das provas e matrizes[77].

Em termos do relato do efeito da AA na avaliação usual de cada escola, foram poucas as justificativas apresentadas. Sete escolas mencionaram a meta-avaliação: “os professores podem avaliar suas formas de avaliar”. Uma escola introduziu novas formas de avaliar e outra usou a AA “como suporte para elaboração das avaliações da Unidade Escolar”. Na formulação da AA, um grande cuidado foi tomado para que suas aplicações não substituíssem as usuais das escolas. Como efeito colateral, entretanto, percebe-se a escola mudando sua forma de avaliar para se adequar ao padrão externo.

Das 290 escolas envolvidas, apenas 19 justificaram a percepção de efeito positivo da AA sobre as práticas pedagógicas. Dentre as justificativas estavam: “aumentou o esforço dos professores na busca de melhoria do desempenho dos alunos / possibilitou sua auto-avaliação” (7), “torna o trabalho eficaz /permite alteração de práticas para melhor atender ao aluno” (5), “oferece ferramentas (matriz/vídeo) que contribuem para repensar/replanejar as práticas pedagógicas” ou ainda a escola “utilizou os materiais da AA” (4) e “foi possível analisar as práticas pedagógicas dos professores que precisam aprimorar-se mais / desenvolvem suas habilidades” (3).

Uma análise das atividades propostas pelas escolas e relatadas no RD das diversas aplicações da AA mostra, entretanto, que boa parte das escolas não associa um objetivo didático a uma determinada atividade. As dificuldades identificadas na 3ª unidade, em Língua Portuguesa, por exemplo, suscitaram atividades tais como: “atividade com uso de dicionário / trabalho com vocabulário” (15), “trabalhar poemas / sarau” (11), trabalhar com parlendas/cantigas de roda/trava-línguas” (8), “estabelecer/manter a hora do conto/reconto” (4), “trabalhar com textos

77 Acresce-se a essas respostas o posicionamento de várias escolas, especialmente em 2001, que relatavam utilizar de maneira lúdica “o caderninho” dos testes para os alunos de 1ª e 2ª série, porque eram cheios de figuras e poderiam ser utilizados em atividades, por exemplo, de “colorir”. Ou ainda, que os alunos gostavam quando ganhavam os cadernos de teste, como se fosse brinde. Embora esses relatos não sejam representativos, a percepção de efeito da AA atrelada ao uso dos cadernos de teste como material didático parece apontar para uma grande carência nas escolas.

instrucionais (receitas e bulas)" (3), "trabalhar quadrinhos/cruzadinhas" (2), "dramatização da síntese de livros literários" (1), ou "concurso / campeonato de leitura" (3). Em Matemática, as propostas foram mais próximas dos descritores. Exemplos das respostas: 148 escolas mencionaram "trabalhar números e operações com problemas do dia a dia", "trabalhar com números e as operações matemáticas básicas" (46), "resgatar o interesse através de jogos educativos" (32), "trabalhar prioritariamente os domínios avaliados" (19), "trabalhar com gráficos e tabelas / conta de luz" (14), "construção de sólidos geométricos / estudo de figuras geométricas" (10), "uso de material dourado" (4), "trabalhar a leitura e interpretação das situações-problema" (3) e "uso do ábaco" (1). A relação entre a dificuldade diagnosticada e a atividade proposta para saná-la não foi, na maioria, estabelecida.

Voltando à possibilidade de as respostas obtidas serem, de alguma maneira, resultantes de uma resposta ritualizada, um outro exemplo reforça essa hipótese: ainda que houvesse uma percepção de que a AA teve um efeito positivo na correção de problemas de aprendizagem apresentados pelos alunos (da 1ª à 4ª série, 95,5% - 94,8% - 57,6% e 96,9%), houve uma queda na freqüência das respostas quando se observa "a aprovação dos alunos para as séries seguintes": 88,6%, 90,7%, 56,6% e 93,4% também da 1ª para a 4ª série. A AA foi formulada a partir do pressuposto de que a correção de problemas identificados ao longo do ano afetaria positivamente a aprovação. Aparentemente, as escolas não fizeram essa relação tão direta e, como, será visto na seção que trata de usos instrumentais, as taxas de aprovação caíram um pouco entre 2001 e 2004.

O que torna as respostas interessantes é o fato de variarem em relação às séries e aos itens. Quando observado o comportamento por série, a 3ª teve a menor freqüência de respostas positivas, quando comparada às demais. Nota-se, por exemplo, que exatamente dessa série foram feitas as maiores críticas ao formato dos testes de Matemática. Os professores o consideraram, ao longo do tempo, "descontextualizado". A 3ª série concentrou os itens de teste na dimensão Número e Operações, como disposto na matriz de referência, e muitos descritores não englobavam a problematização das operações. Esse comportamento da escola poderia indicar, como visto no quadro teórico, que a percepção de utilidade é associada à aceitação dos instrumentos de avaliação, mas a tendência de uma resposta menos positiva da 3ª série não viria a se repetir quando do questionamento feito pelo RD 2004 – 3ª unidade, conforme discutido em breve.

Dentre as áreas da grade, a AA não contribuiu tanto para a comunicação entre coordenadores pedagógicos e professores, quando observada a 3ª série (apenas 19 respostas positivas quando, na

1ª série, foram 208 registros nesse sentido). Quanto a essa comunicação, cabe um esclarecimento: muitas escolas, em particular as menores, não contavam com a presença do coordenador pedagógico. Os próprios dirigentes assumiram esse papel ou a secretaria municipal de educação concentrou um setor pedagógico para apoiar as suas unidades escolares.

Das 20 áreas investigadas em 2003, apenas os itens *Interesse dos pais sobre o desempenho escolar dos alunos* e *Participação dos pais nas atividades escolares* contaram com grande número de respostas sem efeito, negativo ou muito negativo, como visto na tabela a seguir.

Tabela 5: Tabela síntese dos percentuais de respostas negativas, distribuídas nas quatro séries do Fundamental Menor, encaminhadas por 290 escolas nos RD de 2003 – 3ª unidade quanto perguntadas sobre o efeito da AA na sua relação com os pais dos alunos.

	Áreas	Efeito Negativo da Avaliação de Aprendizagem			
		1ª série	2ª série	3ª série	4ª série
1	Interesse dos pais sobre o desempenho escolar dos alunos.	40,7	41,0	36,9	38,6
2	Participação dos pais nas atividades escolares.	42,4	41,4	40,0	40,3

Fonte: Relatório Síntese Avaliação de Aprendizagem 2003 – 3ª unidade (p. 20).

As poucas justificativas apresentadas pelas escolas (6) para a percepção de um efeito negativo ou para a falta de efeito da AA na sua relação com os pais parecem associar o problema não à política, mas aos pais. Foram elas: “os pais participam pouco do desempenho dos filhos”(3); “falta de acompanhamento dos pais torna a atuação do professor difícil, trabalhando sozinho sem conseguir atingir as metas”(2) e “os pais são distantes/desmotivados/desinteressados com o processo de aprendizagem que envolve seus filhos”(1). A culpabilização da família pelas dificuldades dos alunos e pelo seu baixo desempenho foi relatada por este autor em sua dissertação de mestrado (DANTAS, 2005): as escolas faziam seus planos para um “aluno ideal”. O aluno “real” e sua família eram empecilhos para que esses planos fossem concretizados.

B. Percepção em 2004

Para aprofundar o *feedback* das escolas sobre o efeito da AA observado em 2003, novamente na 3ª unidade, em 2004, o RD incluiu um novo bloco de questões. Como a baixa freqüência de resposta às questões no RD 2003-3ª unidade foi associada à utilização da grade com as 20 áreas, dessa vez optou-se por questões diretas e fechadas, para simplificação das respostas. Essa estratégia de simplificação aparentemente foi bem sucedida porque, dentre as 1.200 escolas que encaminharam o RD a tempo do processamento para o Relatório Síntese, um número pequeno (inferior a 20%) não relatou sua percepção, contrário ao que havia ocorrido em 2003.

Da mesma forma que em 2003, é importante registrar que o *feedback* das escolas aqui relatado não representa o conjunto de unidades abrangidas pela AA. Na 3ª unidade de 2004, esse número equivaleu a aproximadamente 42% do universo avaliado. Não há registro sobre os 58% restantes. O interessante seria o delineamento de um estudo de efetividade, com o conjunto de escolas da AA 2001 – 2004, para levantar seus efeitos cinco anos após o fechamento do ciclo e, ao fazê-lo, distinguir as escolas que respeitaram os prazos daquelas que não o fizeram, na busca por diferenciação de comportamento entre os grupos.

Dentre as questões sobre o efeito da AA no RD em 2004, uma disse respeito à percepção do professor sobre a melhoria no aprendizado de Português e Matemática como efeito do trabalho com a AA (e de Produção Textual nas 3ª e 4ª séries). A questão foi dirigida aos professores das quatro séries do Fundamental Menor e teve como alternativas sim, não, e não é possível afirmar. Do grupo de 1.200 escolas cujos RD foram processados, a base foi consolidada com 917 registros válidos para a AA[78]. As três próximas tabelas trazem seu posicionamento.

Tabela 6: Percepção, por parte da escola, de melhoria no aprendizado de Língua Portuguesa nos alunos das 4 séries do Ensino Fundamental Menor em função do trabalho com a Avaliação de Aprendizagem – RD 2004- 3ª unidade.

	Percepção Melhoria Língua Portuguesa	**F**	**%V**	**F**	**%V**	**F**	**%V**	**F**	**%V**
		1ª série		**2ª série**		**3ª série**		**4ª série**	
Válido	Não relata	159	17,3	72	7,9	44	4,8	29	3,2
	Não	39	4,3	39	4,3	47	5,1	29	3,2
	Não é possível afirmar	128	14,0	115	12,5	102	11,1	70	7,6
	Sim	591	64,4	691	75,4	724	79,0	788	86,0
	Total	917	100,0	917	100,0	917	100,0	916	100,
Missing		1.082		1.082		1.082		1.083	
	Total	1.999		1.999		1.999		1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade

Tabela 7: Percepção, por parte da escola, de melhoria no aprendizado de Matemática nos alunos das 4 séries do Ensino Fundamental Menor em função do trabalho com a Avaliação de Aprendizagem – RD 2004 3ª unidade.

	Percepção Melhoria Matemática	**F**	**%V**	**F**	**%V**	**F**	**%V**	**F**	**%V**
		1ª série		**2ª série**		**3ª série**		**4ª série**	
Válido	Não relata	179	19,5	107	11,7	59	6,4	43	4,7
	Não	43	4,7	47	5,1	60	6,5	52	5,7
	Não é possível afirmar	128	14,0	133	14,5	128	14,0	106	11,6
	Sim	567	61,8	629	68,7	670	73,1	715	78,1
	Total	917	100,0	916	100,0	917	100,0	916	100,0
Missing		1.082		1.083		1.082		1.083	
	Total	1.999		1.999		1.999		1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade

78 Na fusão da base AA 2004 3ª unidade com a base AD 2004, no total foram 1.999 escolas.

Tabela 8: Percepção, por parte da escola, de melhoria no aprendizado de Produção Textual nos alunos de 3ª e 4ª séries do Ensino Fundamental Menor em função do trabalho com a Avaliação de Aprendizagem – RD 2004 3ª unidade.

	Percepção Melhoria Produção Textual	**F**	**%V**	**F**	**%V**
		3ª série		**4ª série**	
Válido	Não relata	77	8,4	50	5,5
	Não	102	11,1	63	6,9
	Não é possível afirmar	251	27,4	120	13,1
	Sim	486	53,1	683	74,6
	Total	916	100,0	916	100,0
Missing		1.083		1.083	
	Total	1.999		1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade

De modo geral, a percepção de efeito positivo foi crescente da 1ª para a 4ª série e um pouco mais favorável em Língua Portuguesa que em Matemática (como no ano anterior). Diferente do ano anterior, entretanto, a 3ª série não teve um comportamento discrepante.

Das três áreas avaliadas, o menor número de respostas positivas em termos da percepção do efeito da AA para a melhoria da aprendizagem foi associado à Produção Textual. Essa resposta é interessante porque o trabalho com produção textual só foi introduzido em 2003, na 4ª série, com um formato diferente dos anteriores. A 3ª série foi inserida nessa testagem em 2004, o que talvez explique parcialmente a reação das escolas. Além disso, os relatos das escolas sobre o processo de correção da produção textual (com a introdução de rubricas) o consideraram trabalhoso e, por vezes, difícil.

A resposta positiva menos freqüente na 1ª série pode estar atrelada à percepção, muitas vezes apresentada por professores em fóruns, de que não é possível avaliar alunos tão jovens (média de 07 anos) com instrumentos de múltipla escolha. No caso em relato, as provas eram lidas para os alunos pelos professores em sala de aula. Assim, buscou-se eliminar o efeito do não domínio da leitura dos enunciados das questões, visto que muitas redes não ofereciam pré-escola. Ainda que os testes utilizados tivessem sido construídos com itens pré-testados e que as amostras controladas apontassem para a discriminação entre alunos proficientes x alunos não proficientes, a crítica ao formato dos testes pode estar associada à menor freqüência de respostas positivas para a série, como já discutido para a 3ª série, no ano de 2003.

Comparando-se a percepção das escolas em 2004 com o desempenho dos alunos no estudo monitorado (ver tabela a seguir), não foi possível estabelecer um paralelo entre eles. Por exemplo, a percepção das escolas quanto aos efeitos positivos da AA foi crescente da 1ª série para a 4ª série, mas as maiores defasagens entre o desempenho esperado e o desempenho real não ocorreram nessa seqüência, como pode ser visto na tabela a seguir. Em Matemática, as maiores diferenças entre o esperado e o alcançado foram observadas na 3ª série (- 25,3% em Números e Operações) e na 4ª série (-28,6% em Grandezas e Medidas e -29,9% em Espaço e Forma). A 1ª série, na mesma disciplina, apresentou uma defasagem de 10,6%, bastante inferior à 3ª e à 4ª. Como discutido na Subseção 2.4.1, a percepção do indivíduo é fruto de suas expectativas pessoais e não necessariamente guarda relação com dados concretos.

Tabela 9: Diferença (em pontos percentuais) entre o desempenho médio dos alunos da amostra por domínios/subdomínios/áreas de conteúdo e o percentual mínimo de acertos recomendado (AA 2004 – 3ª unidade).

Teste	Série	Domínio/subdomínio /áreas de conteúdo	Mínimo recomendado	Diferença relativa (em pts percentuais)
Português	1ª série	Pré-leitura	83%	+ 2,4
		Leitura e escrita		-20,5
	2ª série	Leitura	79%	+ 3,8
		Leitura e escrita		-7,6
	3ª série	Idéias essenciais	75%	-5,3
		Recursos lingüísticos		-13,3
		Valor significativo		-6,7
	4ª série	Valor significativo	65%	-15,4
		Idéias essenciais		-18,5
		Recursos lingüísticos		+ 3,1
		Estrutura lingüística		-12,3
Matemática	1ª série	Números e Operações	85%	-10,6
	2ª série	Números e Operações	78%	-23,1
		Grandezas e Medidas		-2,6
		Espaço e forma		+ 10,3
	3ª série	Números e Operações	79%	-25,3
		Grandezas e Medidas		-28,6
	4ª série	Números e Operações	77%	-18,2
		Espaço e forma		-29,9

Fonte: Projeto de Avaliação Externa

Em muitos casos, o discurso das escolas atribuiu o desempenho baixo dos alunos ao estranhamento ao formato de testes, dentre outras razões. Independente dos resultados obtidos, especialmente para a 1ª e a 2ª séries (ou para o CBA1, em 2001), havia uma condenação do

formato de múltipla escolha e, em várias ocasiões, à testagem em si, independente do tipo de instrumento utilizado. Como a literatura sobre avaliação em larga escala costuma identificar a familiaridade com o tipo de instrumento como um elemento que, com o passar do tempo, pode mascarar os resultados, no 4ª ano do ciclo da AA (2004), o RD perguntou à escola qual a percepção sobre o aumento da familiaridade dos alunos com seus testes. A tabela a seguir traz as respostas.

Tabela 10: Percepção, por parte da escola, de aumento de familiaridade dos alunos das 4 séries do Ensino Fundamental Menor com o formato de testes da Avaliação de Aprendizagem – RD 2004 3ª unidade.

Percepção Aumento familiaridade com o formato de teste		**F**	**%V**	**F**	**%V**	**F**	**%V**	**F**	**%V**
		1ª série		**2ª série**		**3ª série**		**4ª série**	
Válido	Não relata	164	17,9	69	7,5	46	5,0	28	3,1
	Não	15	1,6	8	,9	8	,9	4	,4
	Não é possível afirmar	48	5,2	31	3,4	37	4,0	40	4,4
	Sim	690	75,2	809	88,2	825	90,1	845	92,1
	Total	917	100,0	917	100,0	916	100,0	917	100,0
Missing		1.082		1.082		1.083		1.082	
	Total	1.999		1.999		1.999		1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade

A percepção de familiarização dos alunos com o formato de teste foi, como esperado, crescente da 1ª para a 4ª série. É interessante perceber que não houve queda na freqüência de respostas positivas para a 4ª série, ainda que seu aluno fosse convidado a uma mudança no formato. Até a 3ª série, a resposta era marcada na própria prova (o professor as transcreveria posteriormente para a Tabela do Professor). Na 4ª série, era o próprio aluno quem deveria preencher o gabarito[79]. Apesar das diferenças na freqüência da percepção positiva quanto ao aumento da familiaridade dos alunos para o formato de teste, ela foi superior a 75% em todas as séries.

Em síntese, em 2004, do posicionamento das escolas a partir da visão dos seus professores, registrado nas tabelas anteriores, houve uma percepção geral de que a AA contribuiu positivamente para o ensino de Língua Portuguesa e Matemática. Somente no item que endereçou o efeito em Produção Textual 3ª série recebeu 11,1 % de respostas válidas Não; para os demais a freqüência de uma percepção negativa foi inferior a 10%.

79 A inserção do gabarito nos testes da AA para a 4ª série foi uma demanda da própria escola para a equipe central de Avaliação. O argumento apresentado era que o aluno de 4ª série seria submetido à prova da Avaliação de

Pressupôs-se, no presente trabalho, que uma percepção de utilidade sobre a avaliação é uso-conducente, o que levou à proposição do U8 como indicador de qualidade de políticas de avaliação. Considerado o *feedback* de 2003 e de 2004 (U8), percebe-se um panorama favorável à utilização da AA, talvez ainda mais que aquele colocado a partir dos demais itens verificadores dos sete indicadores da categoria Utilidade (U1 a U7). Há dois cuidados, entretanto: 1) esse *feedback* pode estar atrelado a um discurso ritualizado, já que a percepção é mais positiva nas áreas mais enfatizadas pelas peças de comunicação da política de avaliação; 2) o cenário não é representativo do conjunto de escolas envolvidas na AA visto que apenas corresponde ao grupo que encaminhou o RD para a equipe central a tempo de processamento. Por essas duas razões – possível ritualização e não representatividade – o U8 não pode ser analisado como um indicador definitivo.

Na Categoria Utilidade, a AA apresentou elementos uso-conducentes que sugeriam probabilidade de média a alta de concretização do uso. Especialmente nas respostas ao RD em 2004, houve um registro positivo de percepção da utilidade dessa política para um grupo de escolas envolvidas (75% escolas dentre as 42% delas que encaminharam os RD). Mesmo não sendo representativa do conjunto de escolas, houve uma expectativa de uso para a AA. A próxima subseção trata dos registros sobre os usos concretizados.

Desempenho no final do ano e já estaria familiarizado com o formato do teste. Assim, o resultado da AD não ficaria comprometido pelo desconhecimento do aluno sobre como fazer a marcação das respostas.

4.2 Análise da política de Avaliação de Aprendizagem a partir da categoria Uso

4.2.1 Uso Instrumental

Esse trabalho adotou, como conceito de avaliação educacional, a busca de objetivação do julgamento sobre uma determinada realidade, capturada a partir de um recorte entendido como dado relevante, tendo em vista uma tomada de decisões. Da descrição da política apresentada na Subseção 3.1 e das análises dos oito indicadores da categoria U8 (Seção 4.1), percebe-se que a implementação da política de avaliação sob investigação adotou (e divulgou) matrizes de referência (recorte de dados relevantes) e utilizou quadros e guias-diagnóstico para informar a linha de proficiência por turma, em Língua Portuguesa e em Matemática (e, a partir de 2003, em Produção Textual), como base para o julgamento de valor.

Pelas orientações da AA, após aplicação dos testes, correção e preenchimento dos quadros-diagnósticos, os professores deveriam se reunir com a coordenação pedagógica (ou com a direção da escola) para, identificadas as dificuldades de cada turma, propor alternativas, no planejamento da unidade seguinte, com objetivo de saná-las. No início do ano letivo seguinte, caberia a esse grupo de professores planejarem o curso, as unidades e as aulas usando como um dos insumos os diagnósticos obtidos no ano anterior. No cenário da AA, a tomada de decisões esteve atrelada, portanto, ao planejamento de curso, unidade e aulas.

Como visto no Marco Teórico, o uso instrumental clássico de uma avaliação é exatamente aquele que diz respeito à utilização dos resultados na tomada de decisão. No presente trabalho, ainda que se identifique o uso de resultado como um elemento importante para a definição da qualidade da avaliação, o mesmo foi expandido para outros elementos do delineamento avaliativo para além do resultado (matriz, forma de correção, cadernos de teste), como proposto por Weiss e outros (ver Subseção 2.4.1).

Além disso, o trabalho propôs relacionar o uso instrumental à finalidade da política (melhorar a qualidade da educação pública no Estado, “expressa através do sucesso escolar dos alunos”, de acordo com o Manual do PDE, 2001) para discutir o posicionamento de Patton (1988, 1997), para quem o uso instrumental leva imediatamente e diretamente à melhoria da qualidade do objeto sob avaliação. No presente estudo, a melhoria da qualidade do ensino da escola pública foi caracterizada pela média (*theta* – TRI) por escola do desempenho dos alunos da 4ª série em Português e em Matemática na AD 2004.

O primeiro indicador para verificação do uso instrumental é a utilização da AA (resultados e outros elementos) para a tomada de decisões. É sobre esses resultados a subseção a seguir.

4.2.1.1 Uso da AA para a tomada de decisões

Como descrito no Passo V da Subseção 3.4 (capítulo Metodologia), as informações sobre o uso da AA para a tomada de decisões foram obtidas da base de dados construída com as respostas das 1.013 escolas para a Questão 16 do RD 2004 – 1ª unidade. A questão "Marque com um x os materiais que sua escola utilizou para o planejamento de curso em 2004" foi elaborada com 5 alternativas + 1, sendo essa última um campo aberto para que a escola dissesse que elementos outros da AA teria utilizado para o planejamento do ano letivo. A escola pôde marcar mais de uma opção de resposta. Cada alternativa foi, por isso, tratada como uma questão isolada, com duas possibilidades de resposta (sim, marcou a alternativa, e não, deixou a alternativa em branco).

Tabela 11: Respostas das escolas na 1ª unidade de 2004 sobre utilização da AA 2003 para o planejamento de 2004.

Utilização da AA para planejamento em 2004	Não	Sim	Total
Utilizou os diagnósticos das turmas, obtidos nas três unidades de 2003.	442	571	1.013
Utilizou os diagnósticos das turmas, obtidos **apenas** na 3ª unidade de 2003.	942	71	1.013
Utilizou as matrizes de referência de 1ª e 2ª séries.	346	667	1.013
Utilizou as matrizes de referência de 3ª e 4ª séries.	289	724	1.013
Utilizou a matriz de produção textual para a 4ª série.	472	541	1.013

Das 1.013 escolas com dados válidos, 571 (56%) referiram usar os resultados do ano anterior para o planejamento de curso (um pouco mais da metade) e 71 usaram o diagnóstico feito na 3ª unidade. A hipótese que norteou a presente investigação previa que, dentre as contribuições da AA, os resultados seriam elementos pouco utilizados e que outros itens relacionados ao acontecimento da política teriam uso. Segundo Weiss (1998), são vários os elementos da avaliação usados: os achados (resultados), as recomendações (se e quando existentes), as idéias e generalizações, o processo, a discussão. No caso em tela, as escolas referiram o uso de matrizes de referência no planejamento do curso em uma freqüência superior àquela do uso dos resultados. Do grupo respondente, 667 escolas (66%) apontaram o uso das matrizes de 1ª e 2ª séries e 724 (71%) das matrizes de 3ª e 4ª séries. Esse quadro de reação positiva crescente da 1ª para a 4ª série, já observado no U8, se repetiu mais vezes, especialmente nos itens verificadores do uso conceitual.

Ainda sobre materiais da AA utilizados para a tomada de decisão, é interessante perceber o número menor de escolas que utilizaram a matriz de referência de produção textual 4ª série quando comparado ao número de escolas que usaram as matrizes de Português e de Matemática para as 3ª e 4ª séries (724). A matriz de produção textual 4ª série foi introduzida para as escolas no final de 2003. Talvez por falta de acomodação com a matriz ou por julgar o processo mais difícil e trabalhoso que aquele desenvolvido com os testes de Português e Matemática, a escola tenha se valido menos desse documento que dos demais para o seu planejamento de curso.

O resultado sobre o uso da AA para a tomada de decisão precisa ser considerado com cuidado, já que as 1.013 escolas não representam o conjunto de 2.567 unidades envolvidas pela AA na 1ª unidade de 2004. Em especial, por terem entregado o RD a tempo de processamento, é provável terem sido essas unidades as que melhor se relacionaram com a AA. Além disso, é preciso considerar um possível comportamento ritualizado (como já discutido para as respostas de percepção do efeito da AA discutidas na subseção anterior), visto que os documentos de comunicação da AA recomendaram com freqüência a utilização das matrizes, além dos próprios diagnósticos.

Resultados e matrizes foram oferecidos como alternativas de resposta à questão 16 do RD 2004 – 1ª unidade. O último campo dessa pergunta foi aberto e a escola poderia registrar ali outros elementos da AA usados para o planejamento. Poucas foram as unidades que marcaram essa alternativa. Para o planejamento do ano letivo, 16 escolas referiram usar cadernos de teste, 04 falaram sobre uso do manual de correção textual, 07 sobre materiais de apoio dos vídeos didáticos, e 185 os próprios vídeos (um número expressivo, visto que espontâneo). Foram, portanto, as matrizes os elementos mais utilizados para o planejamento de curso.

Diante desse quadro, perguntou-se se haveria alguma relação entre o uso dos resultados e dos outros elementos com o desempenho dos alunos de 4ª série em Matemática e em Português na AD 2004. Para cada uma das cinco alternativas da Q.16, foram comparadas as médias das escolas de desempenho em Português e Matemática dos alunos de 4ª série por dois grupos: o que fez o planejamento com um dos elementos da AA e o que não referiu tê-lo feito. Como observado nas seis tabelas a seguir, não foi possível estabelecer uma associação entre o relato de planejamento e desempenho dos alunos em nenhuma das alternativas. Optou-se por apenas apresentar os resultados da comparação entre os grupos em relação à utilização do diagnóstico feito em 2003 e das matrizes de 3ª e 4ª série.

Tabela 12: Observação das médias em Língua Portuguesa e em Matemática (4ª série – AD2004) das escolas que fizeram o planejamento 2004 com os resultados obtidos na AA 2003 e aquelas que não o fizeram.

Local	Planejamento 2004 com os resultados AA 2003	Nº escolas	Média	Desvio Padrão	Erro Padrão
Língua Portuguesa	Escolas que relatam o planejamento 2004 com os diagnósticos AA nas 3 unidades de 2003	442	-,5698	,71584	,03405
	Escolas que não relatam o planejamento 2004 com os diagnósticos AA nas 3 unidades de 2003	571	-,5577	,82081	,03435
	Total	1.013	-,5630	,77641	,02439
Matemática	Escolas que relatam o planejamento 2004 com os diagnósticos AA nas 3 unidades de 2003	442	-,2420	,78912	,03753
	Escolas que não relatam o planejamento 2004 com os diagnósticos AA nas 3 unidades de 2003	571	-,1950	,82048	,03434
	Total	1.013	-,2155	,80689	,02535

Fonte: Projeto de Avaliação Externa – AA 2004 e AD 2004

Tabela 13: Resultado ANOVA – Observação da média de desempenho da escola em Português 4ª série (AD 2004 *theta* TRI) em dois grupos de escolas: aquelas que fizeram o planejamento de 2004 a partir do diagnóstico das três unidades letivas de 2003 e aquelas que não o fizeram.

ANOVA Português 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	,036	1	,036	,060	,806
Entre grupos	610,006	1011	,603		
Total	610,043	1012			

Tabela 14: Resultado ANOVA – Observação da média de desempenho da escola em Matemática 4ª série (AD 2004 *theta* TRI) em dois grupos de escolas: aquelas que fizeram o planejamento de 2004 a partir do diagnóstico das três unidades letivas de 2003 e aquelas que não o fizeram.

ANOVA Matemática 4ª série	Soma dos quadrados	Df	Mean Square	F	Sig
Dentro dos grupos	,550	1	,550	,844	,358
Entre grupos	658,334	1011	,651		
Total	658,883	1012			

Tabela 15: Observação das médias em Língua Portuguesa e em Matemática (4ª série – AD2004) das escolas que fizeram o planejamento 2004 com as matrizes de referência de 3ª e 4ª séries e aquelas que não o fizeram.

Local	Planejamento 2004 com as matrizes de 3ª e 4ª séries	No escolas	Média	Desvio Padrão	Erro Padrão
Língua Portuguesa	Escolas que relatam o planejamento 2004 com as matrizes AA de 3ª e 4ª séries	289	-,5855	,80822	,04754
	Escolas que não relatam o planejamento 2004 as matrizes AA de 3ª e 4ª séries	724	-,5540	,76373	,02838
	Total	1.013	-,5630	,77641	,02439
Matemática	Escolas que relatam o planejamento 2004 com as matrizes AA de 3ª e 4ª séries	289	-,2600	,90066	,05298
	Escolas que não relatam o planejamento 2004 as matrizes AA de 3ª e 4ª séries	724	-,1977	,76621	,02848
	Total	1.013	-,2155	,80689	,02535

Tabela 16: Resultado ANOVA – Observação da média de desempenho da escola em Português 4ª série (AD 2004 *theta* TRI) em dois grupos de escolas: aquelas que fizeram o planejamento de 2004 a partir das matrizes de 3ª e 4ª séries e aquelas que não o fizeram.

ANOVA Português 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	,205	1	,205	,340	,560
Entre grupos	609,838	1011	,603		
Total	610,043	1012			

Tabela 17: Resultado ANOVA – Observação da média de desempenho da escola em Matemática 4ª série (AD 2004 *theta* TRI) em dois grupos de escolas: aquelas que fizeram o planejamento de 2004 a partir das matrizes de 3ª e 4ª séries e aquelas que não o fizeram.

ANOVA Matemática 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	,802	1	,802	1,232	,267
Entre grupos	658,082	1011	,651		
Total	658,883	1012			

Os usos pretendidos para uma política de avaliação deveriam levar seus beneficiários a atingir a finalidade da política, como discutido na Subseção 2.4.1. A investigação em relato não é conclusiva em relação às possíveis associações entre o uso dos elementos da política no planejamento do ano de 2004 e os resultados dos alunos de 4ª série, obtidos nas provas de Português e Matemática, mas parece indicar que tais associações não existem, pelo menos de maneira significativa.

A finalidade da AA foi a melhoria da qualidade da educação, expressa pela proficiência em Português e em Matemática. Os objetivos específicos da AA estiveram voltados para a redução da reprovação e aumento da aprovação, visto que, idealmente, as escolas sanariam as dificuldades dos alunos durante o ano letivo. Independente da relação com o planejamento, idealmente a política contribuiria para o atingimento dos objetivos e da finalidade maior. A próxima subseção apresenta o comportamento de dois grupos de escolas: as envolvidas por um ciclo completo da AA (ou quase) e aquelas com pouco tempo ou nenhum, em termos das taxas de aprovação, reprovação e abandono. Em muitos casos, essas associações também não foram significativas, como será visto a seguir.

4.2.1.2 Atingimento dos objetivos gerais da política de avaliação de aprendizagem

Como já mencionado, o objetivo divulgado da AA foi proporcionar o diagnóstico de problemas "na aquisição de competências e habilidades pelos alunos, durante o ano letivo, a tempo de serem colocadas em prática ações de remediação que, em último caso, resultariam em uma menor taxa de reprovação e no oferecimento de um melhor serviço educacional". Esse objetivo foi definido em um contexto caracterizado por uma taxa de atendimento aos jovens de 07 a 14 anos superior a 95% e para o qual o problema mais conspícuo era uma defasagem de, por exemplo, 70 % na 5ª a 8ª séries. Tal defasagem era fruto menos da entrada tardia do alunado para a escolarização que, principalmente, das altas taxas de abandono e repetência. O discurso do Educar para Vencer previa que cada escola, na elaboração de seu PDE (Manual do PDE, 2001), diagnosticasse sua situação (inclusive pelas taxas de reprovação, aprovação e abandono/evasão) e definisse metas para melhorar a qualidade de sua oferta.

É importante ressaltar, como discutido por Lipsky (1980), o quão amplos são os objetivos "melhorar a qualidade da oferta" e "oferecimento de um melhor serviço educacional". No presente trabalho, eles foram atrelados ao aumento de proficiência dos alunos da 4ª série em Língua Portuguesa e em Matemática e às variações nas taxas de aprovação (positivas) e de reprovação e abandono (negativas), como definidos nos objetivos específicos da AA.

Inicialmente foi feito um contraste entre as variações nas referidas taxas apresentadas por escolas urbanas e rurais de municípios envolvidos e não envolvidos pela AA. Como discutido na Metodologia (Passo VI, Subseção 3.4), foi calculada a diferença entre a taxa da escola em 2004 e em 2001. Em seguida, foi criada uma nova variável (variação), na qual as diferenças calculadas foram categorizadas como taxa variou como esperado, taxa não variou ou taxa variou de modo não esperado. Por exemplo, no caso das taxas de aprovação, era esperado que Tx 2004 – Tx 2001

fosse um resultado positivo (taxa de aprovação em 2004 superior a de 2001). Uma diferença negativa era não esperada e uma diferença igual a zero foi entendida como não variação da aprovação no período.

Buscou-se identificar alguma tendência de comportamento que caracterizasse aqueles municípios nos quais a AA tivesse atingido um ciclo completo. Dessa forma, seria possível observar o comportamento das escolas diretamente afetadas (urbanas de municípios que se envolveram em um ciclo completo da AA), daquelas indiretamente afetadas (rurais dos mesmos municípios), e de escolas rurais e urbanas localizadas em municípios que não tiveram envolvimento com a política de avaliação. Depois da retirada de Salvador, Nilo Peçanha e Jacobina da base de dados, foram consideradas as taxas de 21.759 escolas públicas rurais e urbanas em 414 municípios da Bahia. As próximas tabelas apresentam a freqüência das escolas pelo tipo de variação das taxas na 1ª série e na 4ª série, considerados a localização e o período de envolvimento com a AA.

Tabela 18: Variação das taxas de aprovação da 1ª série das escolas baianas entre 2004 e 2001, por localização, e o envolvimento de seu município com a AA.

Local da escola	Tempo de envolvimento do município com a AA	Variação da taxa de aprovação da 1ª série						
		Aprovação 1ª série menor em 2004	% do total	Aprovação 1ª série não variou entre 2001 e 2004	% do total	Aprovação 1ªsérie maior em 2004	% do total	Total
Rural	Não participou	285	53,67	48	9,04	198	37,29	531
	Apenas AD 2004	220	44,27	9	1,81	268	53,92	497
	AA 2004	300	53,96	12	2,16	244	43,88	556
	2002-2003	1.035	54,56	51	2,69	811	42,75	1.897
	2001 ou ano anterior	1.000	52,19	60	3,13	856	44,68	1.916
	Total	2.840	52,62	180	3,34	2.377	44,04	5.397
Urbana	Não participou	92	62,16	0	0,00	56	37,84	148
	Apenas AD 2004	75	51,02	1	0,68	71	48,30	147
	AA 2004	84	57,93	0	0,00	61	42,07	145
	2002-2003	213	51,45	2	0,48	199	48,07	414
	2001 ou ano anterior	268	58,52	0	0,00	190	41,48	458
	Total	732	55,79	3	0,23	577	43,98	1.312

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Observando-se a freqüência de ocorrência das escolas nas três categorias criadas (taxa de aprovação em 2004 menor que em 2001 – resultado não esperado, não houve variação da taxa entre os anos, e taxa de aprovação em 2004 maior que em 2001 – resultado esperado), não foi possível estabelecer um padrão em termos do envolvimento ou não envolvimento com a AA para

a 1ª série. A tabela anterior mostra um dado preocupante, no entanto. A tendência tanto na zona rural quanto na urbana foi de uma aprovação menor em 2004.

O mesmo ocorreu com a 4ª série, como pode ser visto na tabela a seguir. Nesse caso, essa ocorrência foi mais forte para as escolas urbanas e para aquelas situadas em municípios com um tempo maior de envolvimento com a AA.

Tabela 19: Variação das taxas de aprovação da 4ª série das escolas baianas entre 2004 e 2001, por localização, e o envolvimento de seu município com a AA.

Local da escola	Tempo de envolvimento do município com a AA	Variação da taxa de aprovação da 4ª série						
		Aprovação 4ª série menor em 2004	% do total	Aprovação 4ª série não variou entre 2001 e 2004	% do total	Aprovação 4ªsérie maior em 2004	% do total	Total
Rural	Não participou	936	43,64	460	21,45	749	34,92	2.145
	Apenas AD 2004	334	48,48	124	18,00	231	33,53	689
	AA 2004	296	50,34	68	11,56	224	38,10	588
	2002-2003	1.805	49,90	519	14,35	1.293	35,75	3.617
	2001 ou ano anterior	1.406	49,61	474	16,73	954	33,66	2.834
	Total	4.777	48,38	1.645	16,66	3.451	34,95	9.873
Urbana	Não participou	226	54,33	3	0,72	187	44,95	416
	Apenas AD 2004	96	50,53	2	1,05	92	48,42	190
	AA 2004	119	58,91	1	0,50	82	40,59	202
	2002-2003	461	59,95	9	1,17	299	38,88	769
	2001 ou ano anterior	525	63,18	6	0,72	300	36,10	831
	Total	1.427	59,26	21	0,87	960	39,87	2.408

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Esse fenômeno já tinha sido levantado quando da avaliação de impacto do Projeto Bahia, conduzida pela FIA/USP em 2003, em termos do efeito da AA. Uma hipótese levantada à época dizia respeito a um movimento em prol de maior rigor nas avaliações internas das escolas. Talvez esse tenha sido o caso para explicar a 4ª série, mas taxas de aprovação mais desfavoráveis em 2004 foram apresentadas por escolas em municípios não integrantes do programa de reforma do governo e foram encontradas também nas escolas rurais (não diretamente afetadas pela AA) dos municípios que fizeram parte do Educar para Vencer.

Quando observadas as escolas (rurais e urbanas) em relação à reprovação na 1ª série, o resultado foi, obviamente, também bastante preocupante: o número de escolas com queda na taxa de reprovação entre 2004 e 2001 foi inferior ao número daquelas que tiveram aumento em suas taxas. O resultado esperado previa que o maior número de escolas tivesse apresentado queda da taxa em

2004 quando comparada à taxa de 2001. O mesmo mecanismo ocorreu para a 4ª série, como pode ser visto nas duas próximas tabelas.

Tabela 20: Variação das taxas de reprovação da 1ª série das escolas baianas entre 2004 e 2001, por localização, e o envolvimento de seu município com a AA.

Local da escola	Tempo de envolvimento do município com a AA	Variação da taxa de reprovação da 1ª série						
		Reprovação 1ª série maior em 2004	% do total	Reprovação 1ª série não variou entre 2001 e 2004	% do total	Reprovação 1ªsérie menor em 2004	% do total	Total
Rural	Não participou	146	52,90	3	1,09	127	46,01	276
	Apenas AD 2004	169	48,42	6	1,72	174	49,86	349
	AA 2004	223	55,06	7	1,73	175	43,21	405
	2002-2003	911	60,77	25	1,67	563	37,56	1.499
	2001 ou ano anterior	715	58,37	21	1,71	489	39,92	1.225
	Total	2.164	57,65	62	1,65	1.528	40,70	3.754
Urbana	Não participou	58	64,44	0	0,00	32	35,56	90
	Apenas AD 2004	59	64,84	0	0,00	32	35,16	91
	AA 2004	79	67,52	0	0,00	38	32,48	117
	2002-2003	195	62,70	1	0,32	115	36,98	311
	2001 ou ano anterior	236	67,62	0	0,00	113	32,38	349
	Total	627	65,45	1	0,10	330	34,45	958

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Tabela 21: Variação das taxas de reprovação da 4ª série das escolas baianas entre 2004 e 2001, por localização, e o envolvimento de seu município com a AA.

Local da escola	Tempo de envolvimento do município com a AA	Variação da taxa de reprovação da 4ª série						
		Reprovação 4ª série maior em 2004	% do total	Reprovação 4ª série não variou entre 2001 e 2004	% do total	Reprovação 4ªsérie menor em 2004	% do total	Total
Rural	Não participou	348	59,90	19	3,27	214	36,83	581
	Apenas AD 2004	116	62,37	6	3,23	64	34,41	186
	AA 2004	115	61,50	1	0,53	71	37,97	187
	2002-2003	643	60,55	33	3,11	386	36,35	1.062
	2001 ou ano anterior	469	57,83	25	3,08	317	39,09	811
	Total	1.691	59,82	84	2,97	1.052	37,21	2.827
Urbana	Não participou	212	59,89	2	0,56	140	39,55	354
	Apenas AD 2004	85	62,04	0	0,00	52	37,96	137
	AA 2004	106	62,35	1	0,59	63	37,06	170
	2002-2003	400	63,49	3	0,48	227	36,03	630
	2001 ou ano anterior	445	66,52	4	0,60	220	32,88	669
	Total	1.248	63,67	10	0,51	702	35,82	1.960

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Nos anos de 2001 a 2004, a variação da taxa de abandono foi a única no sentido esperado: em 2004, o percentual de escolas para as quais a taxa de abandono foi menor que em 2001 superou o percentual do cálculo inverso. Novamente, o comportamento foi observado tanto para escolas rurais como urbanas, como pode ser visto na tabela a seguir.

Tabela 22: Variação das taxas de abandono da 1ª série das escolas baianas entre 2004 e 2001, por localização, e o envolvimento de seu município com a AA.

Local da escola	Tempo de envolvimento do município com a AA	Variação da taxa de abandono da 1ª série						
		Abandono 1ª série maior em 2004	% do total	Abandono 1ª série não variou entre 2001 e 2004	% do total	Abandono 1ªsérie menor em 2004	% do total	Total
Rural	Não participou	126	52,07	5	2,07	111	45,87	242
	Apenas AD 2004	154	51,68	4	1,34	140	46,98	298
	AA 2004	228	57,00	8	2,00	164	41,00	400
	2002-2003	531	46,99	11	0,97	588	52,04	1130
	2001 ou ano anterior	522	46,98	18	1,62	571	51,40	1111
	Total	1.561	49,07	46	1,45	1.574	49,48	3.181
Urbana	Não participou	55	39,57	0	0,00	84	60,43	139
	Apenas AD 2004	56	40,58	1	0,72	81	58,70	138
	AA 2004	58	42,96	1	0,74	76	56,30	135
	2002-2003	134	36,61	0	0,00	232	63,39	366
	2001 ou ano anterior	161	41,28	1	0,26	228	58,46	390
	Total	464	39,73	3	0,26	701	60,02	1.168

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Tabela 23: Variação das taxas de abandono da 4ª série das escolas baianas entre 2004 e 2001, por localização, e o envolvimento de seu município com a AA.

Local da escola	Tempo de envolvimento do município com a AA	Variação da taxa de abandono da 4ª série						
		Abandono 4ª série maior em 2004	% do total	Abandono 4ª série não variou entre 2001 e 2004	% do total	Abandono 4ªsérie menor em 2004	% do total	Total
Rural	Não participou	287	51,43	15	2,69	256	45,88	558
	Apenas AD 2004	88	49,16	8	4,47	83	46,37	179
	AA 2004	139	53,05	2	0,76	121	46,18	262
	2002-2003	578	49,61	33	2,83	554	47,55	1.165
	2001 ou ano anterior	410	49,58	35	4,23	382	46,19	827
	Total	1.502	50,22	93	3,11	1.396	46,67	2.991
Urbana	Não participou	156	51,15	0	0,00	149	48,85	305
	Apenas AD 2004	77	54,61	0	0,00	64	45,39	141
	AA 2004	97	59,88	0	0,00	65	40,12	162
	2002-2003	255	45,86	3	0,54	298	53,60	556
	2001 ou ano anterior	283	49,56	3	0,53	285	49,91	571
	Total	868	50,03	6	0,35	861	49,63	1.735

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Quando observada a 4ª série (tabela anterior), no entanto, o percentual de escolas para as quais as taxas de abandono foram maiores em 2004 supera aquelas nas quais as taxas haviam caído.

O quadro apresentado nas tabelas anteriores não permitiu concluir que houve um movimento de escolas urbanas impactadas pela AA em termos das variações das taxas de aprovação, reprovação e abandono. O próximo passo, então, foi dado no sentido de verificar se havia alguma associação entre a variação das taxas e o período de envolvimento do município com a AA, utilizando-se o Gamma. Como pode ser visto na próxima tabela, não foi encontrada qualquer associação forte ou mediana (Gamma > 0,50). Somente para a 4ª série - urbana foi identificada uma associação significativa fraca (Gamma > 0,10), mas a mesma foi (como já discutido) negativa.

Tabela 24: Resultados de Gamma, ao nível de 95% de confiança, para o cruzamento da variação nas taxas de aprovação, reprovação e abandono da 1ª e 4ª séries do Ensino Fundamental com o ano de envolvimento do município na AA.

Localização	Variação na taxa de **aprovação** da **1ª** série		Variação na taxa de **reprovação** da **1ª** série		Variação na taxa de **abandono** da **1ª** série	
	Gamma	Sig	Gamma	Sig	Gamma	Sig
Rural	-0,004	,854	-0,071	,005	0,075	,004
Urbana	-0,012	,765	-0,033	,514	-0,006	,895
Localização	Variação na taxa de **aprovação** da **4ª** série		Variação na taxa de **reprovação** da **4ª** série		Variação na taxa de **abandono** da **4ª** série	
	Gamma	Sig	Gamma	Sig	Gamma	Sig
Rural	-0,036	,005	0,030	,294	0,013	,630
Urbana	-0,111	,000	-0,077	,027	0,043	,226

Como as associações não se mostraram sequer medianas, foi buscado um segundo caminho de análise, dessa vez por meio da análise da variância (ANOVA). Foram criados dois grandes grupos: municípios que não participaram ou só iniciaram em 2004 (na AA ou mesmo apenas com a Avaliação de Desempenho, AD) e aqueles nos quais a AA cumpriu um ciclo completo ou parcialmente completo. No primeiro grande grupo encontraram-se 34% das escolas urbanas e os 66% restantes, dentre as urbanas, estiveram associadas aos municípios do segundo grupo.

Se observado o envolvimento do município, incluindo escolas rurais e urbanas como integrantes da mesma rede, tem-se um percentual próximo: 33% estão em municípios onde não se esperava um efeito mais direto e 67% em municípios nos quais essa expectativa esteve presente porque teriam cumprido um ciclo inteiro da AA (ou quase). A tabela a seguir apresenta a distribuição das escolas localizadas na zona rural ou urbana e o envolvimento de seus municípios com a AA.

Tabela 25: Freqüência das escolas localizadas em municípios baianos por tempo de envolvimento do município com a política AA

Local	Tempo de envolvimento do município com a AA	Freqüência	% válido
Geral	Nenhum ou até um ano	7.180	33,0
	Maior que 1 ano	14.579	67,0
	Total	**21.759**	**100,0**
Rural	**Tempo de envolvimento do município com a AA**	**Freqüência**	**% válido**
	Nenhum ou até um ano	5.866	32,8
	Maior que 1 ano	12.024	67,2
	Total	17.890	100,0
Urbana	**Tempo de envolvimento do município com a AA**	**Freqüência**	**% válido**
	Nenhum ou até um ano	1.314	34,0
	Maior que 1 ano	2.555	66,0
	Total	3.869	100,0

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Quando observado o comportamento das diferenças nas taxas de aprovação da 1ª série e da 4ª série entre 2004 e 2001, nos grupos rural e urbano, pelo envolvimento de seu município com a AA, não foi possível perceber uma tendência diversa entre os grupos que fosse mediana ou forte e significativa, como pode ser visto nas quatro próximas tabelas.

Tabela 26: Diferenças nas taxas de aprovação da 1ª série (2004 – 2001), por localização, das escolas situadas em municípios envolvidos ou não com a AA.

Local	Tempo de envolvimento do município com a AA –Aprovação 1ª série	Nº escolas	Média	Desvio Padrão	Erro Padrão
Rural	Nenhum ou até um ano	1.584	-2,3668	30,31456	,76168
	Maior que 1 ano	3.813	-3,9322	30,10485	,48753
	Total	5.397	-3,4727	30,17216	,41071
Urbana	Nenhum ou até um ano	440	-4,8765	24,11329	1,14956
	Maior que 1 ano	872	-3,7164	23,65974	,80122
	Total	1.312	-4,1055	23,80992	,65734

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Tabela 27: Resultado ANOVA – diferenças nas taxas de aprovação da 1ª série (2004 – 2001) quando contrastadas com o envolvimento do município com a AA.

Local	ANOVA tx apr 1ª série	Soma dos quadrados	df	Mean Square	F	Sig
Rural	Dentro dos grupos	2742,400	1	2742,400	3,014	,083
	Entre grupos	4909556,633	5395	910,020		
	Total	4912299,033	5396			
Urbana	Dentro dos grupos	393,621	1	393,621	,694	,405
	Entre grupos	742828,194	1310	567,044		
	Total	743221,815	1311			

Tabela 28: Diferenças nas taxas de aprovação da 4ª série (2004-2001), por localização, das escolas situadas em municípios envolvidos ou não com a AA.

Local	Tempo de envolvimento do município com a AA –Aprovação 4ª série	Nº escolas	Média	Desvio Padrão	Erro Padrão
Rural	Nenhum ou até um ano	3.422	-3,5581	28,31805	,48409
	Maior que 1 ano	6.451	-4,3802	27,18746	,33850
	Total	9.873	-4,0952	27,58592	,27763
Urbana	Nenhum ou até um ano	808	-2,9319	21,30756	,74960
	Maior que 1 ano	1.600	-4,4329	20,64258	,51606
	Total	2.408	-3,9292	20,87569	,42541

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Tabela 29: Resultado ANOVA – diferenças nas taxas de aprovação da 4ª série (2004 – 2001) quando contrastadas com o envolvimento do município com a AA.

Local	ANOVA tx apr 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Rural	Dentro dos grupos	1511,193	1	1511,193	1,986	,159
	Entre grupos	7510910,585	9.871	760,907		
	Total	7512421,778	9.872			
Urbana	Dentro dos grupos	1209,515	1	1209,515	2,777	,096
	Entre grupos	1047747,520	2.406	435,473		
	Total	1048957,035	2.407			

Quando observadas as diferenças nas taxas de reprovação da 1ª série, os dois conjuntos (rural e urbano) mostram diferenças significativas em termos do envolvimento dos municípios com a AA, embora apenas sejam grandes nas escolas rurais. Esses dados merecem uma investigação posterior. As duas tabelas a seguir apresentam os resultados da 1ª série. Vale a pena ressaltar que as médias das diferenças 2004 – 2001 são grandes e positivas, quando o resultado esperado – em termos de reprovação – deveria ter sido negativo.

Tabela 30: Diferenças nas taxas de reprovação da 1ª série (2004 – 2001), por localização, das escolas situadas em municípios envolvidos ou não com a AA.

Local	Tempo de envolvimento do município com a AA – Reprovação 1ª série	Nº escolas	Média	Desvio Padrão	Erro Padrão
Rural	Nenhum ou até um ano	1.030	1,9724	27,10204	,84447
	Maior que 1 ano	2.724	6,0441	26,29929	,50390
	Total	3.754	4,9270	26,58051	,43383
Urbana	Nenhum ou até um ano	298	8,1272	19,44856	1,12663
	Maior que 1 ano	660	5,6760	14,98362	,58324
	Total	958	6,4385	16,53103	,53409

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Tabela 31: Resultado ANOVA – diferenças nas taxas de reprovação da 1ª série (2004 – 2001) quando contrastadas com o envolvimento do município com a AA.

Local	ANOVA tx repr 1ª série	Soma dos quadrados	df	Mean Square	F	Sig
Rural	Dentro dos grupos	12390,852	1	12390,852	17,615	,000
	Entre grupos	2639191,552	3.752	703,409		
	Total	2651582,405	3.753			
Urbana	Dentro dos grupos	1233,503	1	1233,503	4,530	,034
	Entre grupos	260290,657	956	272,271		
	Total	261524,160	957			

Quando observado o comportamento das escolas quanto às diferenças entre as taxas de reprovação de 2004 e 2001, volta-se ao achado em termos de taxas de aprovação: não houve uma tendência de comportamento, seja rural ou urbano, que possa ser associado ao envolvimento do município com a AA. Comparadas às diferenças da reprovação na 1ª série, os achados para a 4ª série são menores, ainda que também positivos, como mostra a próxima tabela.

Tabela 32: Diferenças nas taxas de reprovação da 4ª série 2004 - 2001, por localização, das escolas situadas em municípios envolvidos ou não com a AA.

Local	Tempo de envolvimento do município com a AA – Reprovação 4ª série	Nº escolas	Média	Desvio Padrão	Erro Padrão
Rural	Nenhum ou até um ano	954	5,7580	23,24174	,75248
	Maior que 1 ano	1.873	4,2429	21,75029	,50257
	Total	2.827	4,7542	22,27218	,41889
Urbana	Nenhum ou até um ano	661	2,9818	14,39071	,55973
	Maior que 1 ano	1.299	4,2494	14,02633	,38917
	Total	1.960	3,8219	14,15928	,31983

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Tabela 33: Resultado ANOVA – diferenças nas taxas de reprovação da 4ª série (2004 – 2001) quando contrastadas com o envolvimento do município com a AA.

Local	ANOVA tx repr 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Rural	Dentro dos grupos	1451,048	1	1451,048	2,927	,087
	Entre grupos	1400386,822	2825	495,712		
	Total	1401837,870	2826			
Urbana	Dentro dos grupos	703,862	1	703,862	3,515	,061
	Entre grupos	392046,780	1958	200,228		
	Total	392750,642	1959			

O mesmo comportamento observado nas diferenças das taxas de reprovação da 1ª série foi notado nas diferenças das taxas de abandono daquela série. Interessantemente, os sinais foram negativos

na maior parte dos grupos, como se esperava que fossem (pelas expectativas dos formuladores das políticas), especialmente nas escolas da zona urbana. Novamente, a maior diferença entre grupos se deu na zona rural, que passou de taxa positiva para negativa quando observados os grupos de escola por envolvimento do município com a AA, enquanto a zona urbana não apresentou variação forte ou significativa nesses grupos.

Tabela 34: Diferenças nas taxas de abandono da 1ª série (2004 – 2001), por localização, das escolas situadas em municípios envolvidos ou não com a AA.

Local	Tempo de envolvimento do município com a AA – Abandono 1ª série	Nº escolas	Média	Desvio Padrão	Erro Padrão
Rural	Nenhum ou até um ano	940	2,5827	26,93416	,87850
	Maior que 1 ano	2.241	-1,4300	24,65746	,52087
	Total	3.181	-,2442	25,41330	,45059
Urbana	Nenhum ou até um ano	412	-6,0644	24,48948	1,20651
	Maior que 1 ano	756	-6,0104	20,40228	,74202
	Total	1.168	-6,0294	21,92069	,64141

Fonte: SEC/MEC Censo Escolar 2001 e 2004

Tabela 35: Resultado ANOVA – diferenças nas taxas de abandono da 1ª série (2004 – 2001) quando contrastadas com o envolvimento do município com a AA.

Local	ANOVA tx abandono 1ª série	Soma dos quadrados	df	Mean Square	F	Sig
Rural	Dentro dos grupos	10663,158	1	10663,158	16,592	,000
	Entre grupos	2043095,233	3179	642,685		
	Total	2053758,391	3180			
Urbana	Dentro dos grupos	,779	1	,779	,002	,968
	Entre grupos	560762,143	1166	480,928		
	Total	560762,922	1167			

Na 4ª série, a diferença média entre as taxas de abandono 2004 – 2001 conservou o sinal negativo (à exceção das escolas rurais nos municípios onde não houve a implementação da AA ou o envolvimento foi pequeno), mas o valor foi inferior àquele percebido na 1ª série. Observa-se que tanto para as escolas urbanas quanto para as rurais, os resultados são mais favoráveis (taxa negativa) nos grupos cujo envolvimento com a AA foi superior a 1 ano, como pode ser visto na próxima tabela.

Tabela 36: Diferenças nas taxas de abandono da 4ª série (2004 – 2001), por localização, das escolas situadas em municípios envolvidos ou não com a AA.

Local	Tempo de envolvimento do município com a AA – Abandono 4ª série	Nº escolas	Média	Desvio Padrão	Erro Padrão
Rural	Nenhum ou até um ano	999	,5772	26,90956	,85138
	Maior que 1 ano	1.992	-,3525	24,24229	,54316
	Total	2.991	-,0420	25,16394	,46012
Urbana	Nenhum ou até um ano	608	-,1551	18,81834	,76318
	Maior que 1 ano	1.127	-2,1381	18,37764	,54743
	Total	1.735	-1,4432	18,55200	,44539

Fonte: SEC/MEC Censo Escolar 2001 e 2004

De qualquer maneira, a variação da diferença da taxa de abandono das escolas urbanas foi significativa entre os grupos, mas pequena, como mostram as duas tabelas a seguir.

Tabela 37: Resultado ANOVA – diferenças nas taxas de abandono da 4ª série (2004 – 2001) quando contrastadas com o envolvimento do município com a AA.

Local	ANOVA tx abandono 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Rural	Dentro dos grupos	575,090	1	575,090	,908	,341
	Entre grupos	1892764,594	2.989	633,243		
	Total	1893339,683	2.990			
Urbana	Dentro dos grupos	1552,973	1	1552,973	4,521	,034
	Entre grupos	595249,604	1.733	343,479		
	Total	596802,578	1.734			

De maneira geral, em relação às variações de taxas de aprovação, reprovação e abandono das escolas, quando associadas aos períodos de entrada de seus municípios na AA, não houve qualquer tendência que apontasse para o resultado esperado, à exceção das taxas de abandono em alguns grupos. Na verdade, houve uma tendência geral contrária de aumento de reprovação e diminuição da aprovação, não necessariamente relacionada à entrada do município na AA. Como já mencionado, uma das hipóteses levantadas para essa tendência é que as escolas estariam mais rigorosas na busca pela qualidade de ensino e evitando aprovação de alunos que ainda não tivessem o domínio dos conteúdos. O dado bom é que, a ser comprovado esse comportamento, o mesmo não resultou em abandono da escola por parte do aluno.

Voltando-se para os objetivos geral e específicos da AA, além da expectativa de efeito – não observado – para o aumento de aprovação e queda da reprovação, esperava-se que o desempenho do alunado nas disciplinas básicas (Português e Matemática) fosse melhorado. Por essa razão, foi interessante analisar o comportamento do desempenho médio das escolas quando observado o

período de envolvimento do seu município com a AA. Diferente das análises anteriores, que puderam ser feitas com escolas urbanas e rurais e com municípios envolvidos ou não com a AA, no caso do desempenho só foi possível observar o comportamento das escolas envolvidas pela AD (todas urbanas). O grupo de contraste (escolas em municípios nunca envolvidos pela AA ou mesmo escolas rurais) foi reduzido a 37 unidades, como pode ser visto na distribuição apresentada pela próxima tabela.

Tabela 38: No de escolas com dados válidos – Desempenho em Língua Portuguesa e em Matemática em *theta* TRI AD 2004

Tempo de envolvimento do município com a AA – AD Português e Matemática 4ª série 2004	Nº escolas
AD 2004	37
AA 2004	205
2002-2003	866
2001 ou anos anteriores	891
Total	1.999

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Os próximos gráficos ilustram as diferenças entre as médias de *theta* (TRI) dos alunos de 4ª série das 1.999 com dados da AD 2004 válidos.

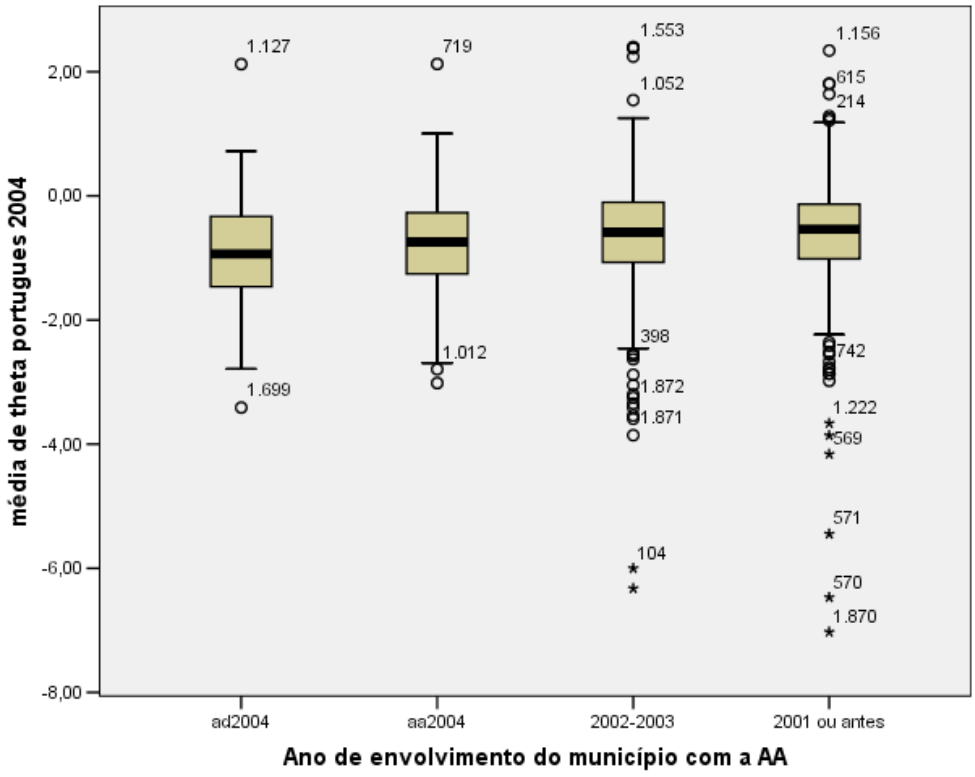


Ilustração 19: *Boxplot* Desempenho 4ª série Português em *theta* – TRI AD 2004 x Ano de envolvimento do município com a AA

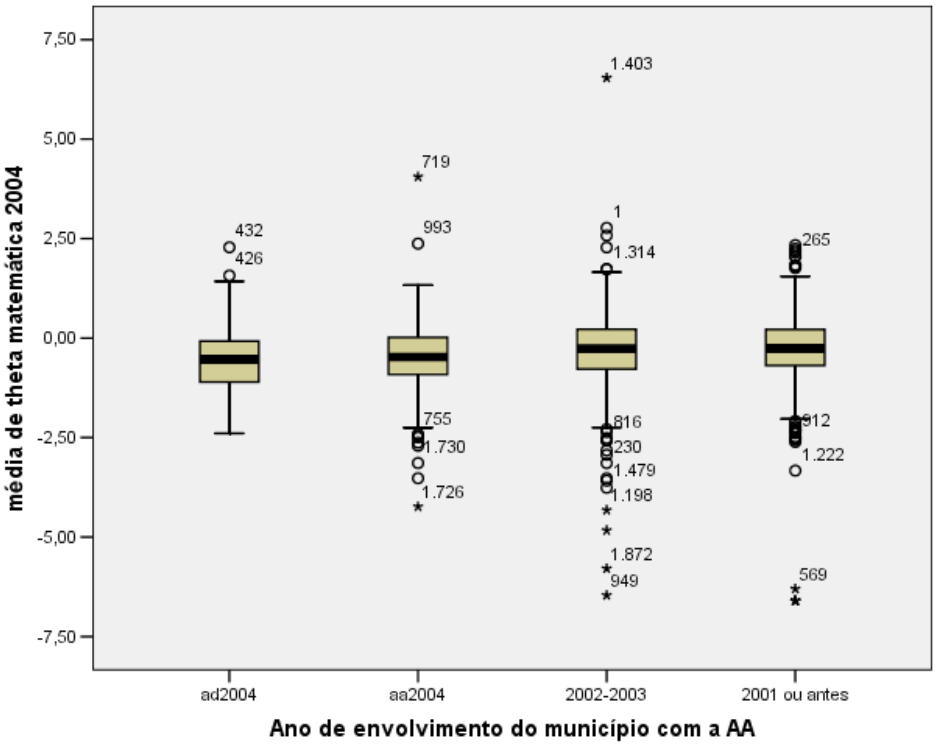


Ilustração 20: *Boxplot* Desempenho 4ª série Matemática em *theta* – TRI AD 2004 x Ano de envolvimento do município com a AA.

Como os gráficos mostraram que, pelo menos visualmente, havia alguma variação nas médias de resultados das escolas situadas nos municípios com diferentes períodos de envolvimento com a AA, optou-se pelo agrupamento das escolas em dois diferentes conjuntos: aquelas situadas em municípios com mais de um ano de envolvimento e aquelas situadas em municípios com menos de um ano ou nenhum envolvimento. A análise de variância das médias de desempenho mostrou resultados significativos, que podem ser vistos nas tabelas a seguir.

Tabela 39: Desempenho em Língua Portuguesa (*theta* TRI) na Avaliação de Desempenho em 2004 por envolvimento de seus municípios na AA.

Tempo de envolvimento do município com a AA – AD Português 4ª série	Freqüência	Média	Desvio Padrão	Erro Padrão
Nenhum ou até um ano	245	-,8194	,82920	,05298
Maior que 1 ano	1754	-,6204	,80191	,01915
Total	1999	-,6448	,80773	,01807

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Vale lembrar que o valor de *theta* varia rotineiramente entre + 4 a – 4 e que *thetas* negativos indicam proficiência baixa. Observa-se que a proficiência encontrada nas escolas situadas em

municípios com envolvimento superior a um ano foi mais alta que aquela das escolas situadas em municípios com nenhum ou até um ano de envolvimento.

Tabela 40: Resultado ANOVA – Desempenho da escola em Língua Portuguesa 4ª série (AD 2004 *Theta* TRI) x envolvimento do seu município com a AA.

ANOVA Português 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	8,506	1	8,506	**13,116**	**,000**
Entre grupos	1295,049	1997	,648		
Total	1303,555	1998			

O mesmo comportamento percebido em Português foi encontrado para Matemática e este foi igualmente significativo, como pode ser visto nas duas tabelas que se seguem.

Tabela 41: Desempenho em Matemática (*theta* TRI) na Avaliação de Desempenho em 2004 por envolvimento de seus municípios na AA.

Tempo de envolvimento do município com a AA – AD Matemática 4ª série	Freqüência	Média	Desvio Padrão	Erro Padrão
Nenhum ou até um ano	245	-,5232	,95124	,06077
Maior que 1 ano	1754	-,2932	,88423	,02111
Total	1999	-,3214	,89565	,02003

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Tabela 42: Resultado ANOVA – Desempenho da escola em Matemática 4ª série (AD 2004 *theta* TRI) x envolvimento do seu município com a AA.

ANOVA Matemática 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	11,372	1	11,372	**14,271**	**,000**
Entre grupos	1591,388	1997	,797		
Total	1602,760	1998			

Esse resultado, como os demais, deve ser analisado com cautela. Em primeiro lugar, há um relato claro das escolas sobre a acomodação dos alunos com o formato das provas, o que pode ter tido um efeito real no seu desempenho. Há também o fato de que a análise do envolvimento com a política da avaliação de aprendizagem foi feita pela sua localização (se rural ou urbana) com o município. É possível que várias escolas urbanas em um determinado município parceiro do Educar para Vencer, desde o início e por algum problema relacionado ao registro no censo ou logístico de entrega de materiais, tenham sido deixadas de fora de uma ou mais aplicações de testes da AA ao longo do tempo.

Na discussão do efeito da AA para a melhoria do desempenho dos alunos nos testes de Português e de Matemática, há um outro dado interessante: o grupo de escolas que encaminharam o RD a tempo de seu processamento, na 3ª unidade de 2004, teve seus resultados médios superiores àquelas escolas que não o fizeram. Para essa análise, foram retirados da base AD 2004 as escolas que não tiveram qualquer envolvimento com a AA. As tabelas a seguir apresentam os resultados da comparação entre os dois grupos.

Tabela 43: Desempenho em Língua Portuguesa (*theta* TRI) na Avaliação de Desempenho em 2004 por encaminhamento do RD na 3ª unidade de 2004 – AA.

Tempo de envolvimento do município com a AA – AD Português 4ª série	Freqüência	Média	Desvio Padrão	Erro Padrão
Escolas que não encaminharam RD	1.045	-,7003	,81942	,02535
Escolas que encaminharam RD	917	-,5714	,77737	,02567
Total	1.962	-,6401	,80243	,01812

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Tabela 44: Resultado ANOVA – Desempenho da escola em Língua Portuguesa 4ª série (AD 2004 *Theta* TRI) x por encaminhamento do RD na 3ª unidade de 2004 – AA.

ANOVA Português 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	8,126	1	8,126	12,696	,000
Entre grupos	1254,545	1.960	,640		
Total	1262,671	1.961			

Tabela 45: Desempenho em Matemática (*theta* TRI) na Avaliação de Desempenho em 2004 por encaminhamento do RD na 3ª unidade de 2004 – AA.

Tempo de envolvimento do município com a AA – AD Matemática 4ª série	Freqüência	Média	Desvio Padrão	Erro Padrão
Escolas que não encaminharam RD	1.045	-,4128	,96808	,02995
Escolas que encaminharam RD	917	-,2097	,78467	,02591
Total	1.962	-,3179	,89265	,02015

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Tabela 46: Resultado ANOVA – Desempenho da escola em Matemática 4ª série (AD 2004 *theta* TRI) x por encaminhamento do RD na 3ª unidade de 2004 – AA.

ANOVA Matemática 4ª série	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	20,156	1	20,156	25,613	,000
Entre grupos	1542,409	1.960	,787		
Total	1562,565	1.961			

Novamente os resultados são insatisfatórios, por serem negativos, mas a proficiência é significativamente maior nas escolas que encaminharam os RD quando comparadas àquelas que não enviaram os documentos. Assim como na análise anterior, os dados devem ser olhados com cautela já que podem ser fruto de: 1) acomodação com o formato da avaliação e 2) ritualização, ao mesmo tempo.

Nessa subseção, foram apresentados os dados que relacionam as contribuições da AA para o atingimento do objetivo maior (finalidade) e dos objetivos específicos da política de avaliação do Estado. Embora as taxas de aprovação e reprovação não tenham apresentado o comportamento esperado, houve queda na taxa de abandono tanto na 1ª quanto na 4ª série entre 2004 e 2001. Entretanto, não se pode afirmar que as mudanças nas taxas de aprovação, reprovação e abandono tenham tido relação com o envolvimento do município com a AA. Já quando observados os desempenhos dos alunos de 4ª série em Português e em Matemática, no entanto, houve uma diferença significativa entre as médias das escolas situadas em municípios que se envolveram com a AA em 2003 ou anos anteriores quando comparadas àquelas cujos municípios apenas entraram em 2004 ou simplesmente não tiveram contato com a política. Do mesmo modo, houve uma diferença significativa entre as médias das escolas em Português e especialmente em Matemática quando observado o encaminhamento do RD em resposta à aplicação na 3ª unidade de 2004. Nos dois casos, com resultados interessantes e esperados, não se pode dizer, no entanto, que sejam decorrentes da AA.

A dificuldade na busca do uso instrumental e do estabelecimento de uma relação entre o uso e a finalidade permanece no levantamento do uso conceitual, que se dá quando os usuários não têm condições de utilizar instrumentalmente os achados, mas tais resultados mudam sua percepção sobre o programa e seus efeitos. A próxima subseção apresenta os relatos de uso conceitual, compondo assim o panorama das contribuições da AA na categoria Uso.

4.2.2 Uso Conceitual

Nessa pesquisa, o Uso Conceitual foi discutido a partir de três dimensões: o uso político-persuasório, que trata das estratégias para obtenção de apoio para a mudança de elementos, no contexto avaliado, sobre os quais a escola não tem autonomia, ou para envolvimento da comunidade, ou ainda para o exercício de algum tipo de pressão; o uso motivacional, no qual o simples fato de estar sob avaliação pode implicar, positiva ou negativamente, o indivíduo; e o uso

de partilha, no qual os elementos da avaliação são utilizados para o compartilhamento de uma visão da realidade de ensino onde se atua.

4.2.2.1 Uso político-persuasório da AA

Para a definição do primeiro item verificador do uso conceitual político-persuasório, levantou-se a hipótese de que, à medida que a escola avaliasse a si própria e refletisse sobre sua realidade ao longo do ciclo da AA, tornar-se-ia mais crítica consigo mesma, o que a ajudaria a buscar apoios para os elementos de mudança. Nesse sentido, quanto mais tempo o município no qual a escola estivesse inserida fosse envolvido pela AA, tanto mais acentuado seria esse traço crítico. Uma das manifestações seria identificar pontos fracos para os quais deveria mudar. O item verificador utilizado disse respeito à identificação das lacunas nas competências dos professores que levassem a escola por demanda por capacitação e posterior solicitação, aos órgãos centrais, de apoio direcionado para seu atendimento.

Dados os limites das informações coletadas pelos RDs, no presente estudo, foram focalizadas as demandas para capacitação em Língua Portuguesa e em Matemática, disciplinas avaliadas tanto pela AA quanto pela AD. Como pode ser observado nas tabelas a seguir, dentre as 917 escolas com dados válidos na base AA 2004[80] – 3ª unidade, 48,1% declararam necessidade de capacitação em Língua Portuguesa e 36,7% em Matemática.

Tabela 47: Relato, por parte das escolas envolvidas pela AA, da necessidade de capacitação docente em Português – RD 2004 – 3ª unidade

Necessidade de capacitação em Português		Nº de escolas	% válido
Válido	Não relata	476	51,9
	Sim	441	48,1
	Total	917	100,0
Missing		1.082	
	Total	1.999	

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

80 A base de dados original foi construída com 1.200 dados (escolas) válidos; após cruzamento com os dados da AD 2004, para comparação de médias, esse número caiu para 917.

Tabela 48: Relato, por parte das escolas envolvidas pela AA, da necessidade de capacitação docente em Matemática– RD 2004 – 3ª unidade.

Necessidade de capacitação em Matemática		Nº de escolas	% válido
Válido	Não relata	580	63,2
	Sim	337	36,8
	Total	917	100,0
Missing		1.082	
	Total	1.999	

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Esses dados são bastante interessantes porque não guardaram relação com os próprios diagnósticos feitos por essas mesmas escolas, quando observado o desempenho do alunado. Rotineiramente, ao longo dos quatro anos do ciclo da AA, os resultados em Matemática foram ainda mais preocupantes que aqueles em Português. Entretanto, a escola identificou lacunas em Português em maior freqüência que em Matemática. É possível que essa identificação seja reflexo do distanciamento que a comunidade escolar coloca entre suas práticas e o resultado obtido. O distanciamento é fruto da responsabilização do aluno (e de sua família) por seu (baixo) desempenho (DANTAS, 2005) e também (como visto anteriormente) pelo planejamento didático (em especial das atividades em sala de aula) sem uma definição de objetivos ou um atrelamento ao diagnóstico feito.

Para observar se o tempo de envolvimento com a AA teria uma relação com o desenvolvimento do traço crítico, foi feito cruzamento entre a demanda de capacitação e tempo de envolvimento com a AA. As associações analisadas foram fraquíssimas e não significativas (Gamma). Na seqüência, buscou-se investigar se a mudança de atitude ou percepção para com a capacitação docente teria tido efeito sobre o desempenho do aluno de 4ª série em Português e Matemática. Em outras palavras, se um uso conceitual da avaliação (a mudança de percepção ou de postura crítica) teria afetado o desempenho do aluno. Optou-se pela criação de dois grupos – um, mais crítico, representado pelas escolas que identificaram necessidade de capacitação e o outro, com as unidades escolares que não o fizeram – e por comparar seus resultados em Português e em Matemática na AD 2004 (médias de *theta* – TRI por escola). Os resultados ANOVA não foram significativos.

Em que pese o fato de um percentual alto de escolas ter identificado necessidade de capacitação para seu professorado, não foram encontrados relatos de uso dos resultados para pressionar os

órgãos centrais por apoio. A leitura dos relatórios síntese do monitoramento da AA ao longo do ciclo aponta para o encaminhamento de todas as demandas ao Projeto de Avaliação Externa, fossem relacionadas à capacitação, à necessidade de novos materiais ou de apoio para mudanças mais estruturais. Não foram encontrados registros de que as escolas tivessem utilizados esses dados para solicitar a seus órgãos centrais tais capacitações ou materiais, tendo usado como argumento os resultados obtidos na AA. Como já mencionado no contexto (Subseção 3.1), a política da avaliação foi percebida de maneira isolada, sem articulação com outros programas públicos educacionais.

O segundo item verificador da categoria Uso Conceitual, na dimensão político-persuasório, referiu-se à utilização de elementos da avaliação para o envolvimento da comunidade, nesse estudo representada pelos pais dos alunos. Como visto anteriormente na seção 4.1.9 (percepção de utilidade da AA para a escola), os aspectos da relação pais x escola foram aqueles que mais trouxeram um relato negativo por parte da equipe escolar na 3ª unidade de 2003. Em 2004, foi perguntado às escolas se haviam usado os materiais da AA para o envolvimento dos pais. A tabela a seguir traz a freqüência das respostas obtidas.

Tabela 49: Relato, por parte das escolas, do uso dos materiais da AA para envolvimento dos pais dos alunos– RD 2004 – 3ª unidade.

Uso da AA para envolvimento dos pais		**Nº de escolas**	**% válido**
Válido	Não relata	22	2,4
	Não	322	35,2
	Sim	571	62,4
	Total	915	100,0
Missing		1.084	
	Total	1.999	

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Apesar do posicionamento negativo demonstrado em 2003, 62,4% das escolas que encaminharam os RD a tempo de processamento em 2004 – 3ª unidade referiram utilização dos materiais (não só dos resultados) da AA para o envolvimento de pais de alunos. Na investigação da relação entre o cumprimento do ciclo da política e o uso para envolvimento dos pais, não foi encontrada uma associação forte ou significativa (Gamma). Comparando-se as médias *theta* das escolas, obtidas dos resultados de seus alunos de 4ª série em Português e Matemática, pelos dois grupos (escola relata que a AA contribuiu para o envolvimento com os pais x AA não contribuiu), a diferença encontrada entre os grupos foi muito pequena e não significativa (ANOVA).

O último item de verificação da categoria Uso Conceitual na sua dimensão político-persuasória foi o uso de materiais da AA para que a direção da escola monitorasse (controlasse) o trabalho dos professores. Quando observado o discurso oficial, a AA deveria ser utilizada para promoção de uma discussão coletiva sobre os problemas identificados e um acerto, também coletivo, sobre ações que deveriam ser implementadas para saná-los. Entretanto, nos RD anteriores a 2004, houve relatos espontâneos da utilização da AA, por parte da coordenação pedagógica ou da direção da escola, como elemento de acompanhamento (como controle e regulação) do trabalho docente. O RD 2004-3ª unidade incluiu duas questões sobre esse tema: a primeira voltada para a orientação do trabalho docente no planejamento do curso e a segunda, do seu acompanhamento. As tabelas a seguir mostram a freqüência das respostas pelas alternativas oferecidas.

Tabela 50: Relato, por parte das escolas envolvidas pela AA, de sua contribuição para a orientação dos professores no planejamento do curso – RD 2004 – 3ª unidade.

Efeito da AA na orientação docente no planejamento		**Nº de escolas**	**% válido**
Válidos	Não relata	134	14,6
	Nada	33	3,6
	Pouco	110	12,0
	Suficiente	415	45,3
	Muito	225	24,5
	Total	917	100,0
Missing		1.082	
	Total	1.999	

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Tabela 51: Relato, por parte das escolas envolvidas pela AA, da sua contribuição para o monitoramento dos professores – RD 2004 – 3ª unidade.

Efeito da AA no monitoramento dos professores		**Nº de escolas**	**% válido**
Válidos	Não relata	161	17,6
	Nada	43	4,7
	Pouco	107	11,7
	Suficiente	378	41,2
	Muito	228	24,9
	Total	917	100,0
Missing		1.082	
	Total	1.999	

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Quase 70% das respostas indicaram contribuição da AA no planejamento docente, enquanto 66,1 % das respostas mostram que houve também contribuição no sentido do monitoramento do trabalho docente. Enquanto a primeira pergunta poderia ser interpretada no contexto do planejamento geral, uso pretendido para a AA, a segunda se apresentou em um contexto de uso não previsto. Usar os resultados da AA para monitorar / controlar a equipe docente sugere, como visto anteriormente na Subseção 2.4.1, um desvio de uso em termos daquilo proposto pelos formuladores da AA, sem necessariamente constituir má fé. Sugere também, como demandado principalmente das secretarias municipais de educação, que houve uma demanda real por instrumentos que facilitassem o acompanhamento e o controle, bastante diferente do desenho da AA, que resultou na oferta de instrumentos para replanejamento participativo.

Quando observado o comportamento das escolas agrupadas por ano de envolvimento do município com a AA (Gamma), mais uma vez, as associações foram fraquíssimas e não significativas. Não houve uma tendência de uma maior ou menor utilização da AA para o monitoramento do trabalho docente que pudesse ser associada ao cumprimento de um ciclo completo da política. Após reagrupamento das respostas em dois grupos (escolas que consideraram que a AA contribuiu para o monitoramento e escolas que não referiram esse efeito), foi feita comparação das médias dos *theta*-TRI em Português e Matemática (ANOVA), mas, assim como nos casos anteriores, os resultados não foram significativos.

4.2.2.2 Uso motivacional

A próxima dimensão, na categoria Uso Conceitual, passível de verificação no contexto da AA disse respeito ao efeito motivacional da avaliação para o aluno. Uma das questões levantadas na negociação durante a formulação da AA era relativa ao temor de que testes padronizados, produzidos externamente, pudessem suscitar uma reação negativa – ou de medo - nos alunos. Surpreendentemente, desde os primeiros relatos no RD, ainda em 2001, houve o registro do oposto: as escolas tinham a percepção de que seus alunos estavam mais motivados a aprender em conseqüência da AA e faziam esse registro de maneira espontânea. Na 3ª unidade de 2004, essa questão foi sistematizada no RD, por meio de uma pergunta direcionada aos professores e repetida nas quatro séries. A tabela a seguir traz as respostas válidas.

Tabela 52: Percepção, por parte da escola, do aumento no interesse de aprender dos alunos das 4 séries do Ensino Fundamental Menor em função do trabalho com a Avaliação de Aprendizagem – RD 2004 3ª unidade.

Percepção Aumento do Interesse em Aprender		**F**	**%V**	**F**	**%V**	**F**	**%V**	**F**	**%V**
		1ª série		**2ª série**		**3ª série**		**4ª série**	
Válido	Não relata	171	18,6	74	8,1	50	5,5	36	3,9
	Não	50	5,5	57	6,2	67	7,3	64	7,0
	Não é possível afirmar	149	16,2	128	14,0	137	15,0	102	11,1
	Sim	547	59,7	658	71,8	662	72,3	714	77,9
	Total	917	100,0	917	100,0	916	100,0	916	100,0
Missing		1.082		1.082		1.083		1.083	
	Total	1.999		1.999		1.999		1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade

O padrão das respostas *Sim* mostra que, da 1ª para a 4ª série, houve uma tendência de aumento na percepção de que a AA exerceu um efeito motivacional para o aluno. Além disso, houve uma queda no percentual de escolas que não relataram efeito. Na contrapartida, é interessante perceber que o percentual de respostas *Não* foi maior na 3ª série, seguida da 4ª série. Essas duas séries acumularam os diagnósticos mais preocupantes na mesma unidade letiva.

Não foi possível estabelecer uma associação entre o tempo de exposição do município à AA e a percepção do efeito da política como motivacional. Após recodificação da variável (AA contribuiu para o aumento do interesse x não contribuiu), foi feita uma comparação de médias *theta* -TRI entre os grupos (ANOVA). Como nos casos anteriores, não houve variação significativa entre os grupos. Em outras palavras, a percepção do professor do aumento do interesse em aprender do aluno não necessariamente guardou relação com seu desempenho em Português ou Matemática.

Observado o uso motivacional da AA para os alunos, a próxima subseção discute o uso de partilha.

4.2.2.3 Uso de partilha

A última dimensão investigada na categoria Uso Conceitual foi aquela que tratou dos elementos da avaliação como facilitadores da construção de uma visão compartilhada sobre um determinado diagnóstico ou plano. Essa visão do uso conceitual como elemento de compartilhamento tem sido relatada como ponto positivo em diversas abordagens de avaliação, como o Marco Lógico mencionado na Fundamentação Teórica. Especificamente, no presente estudo, foram observados:

a ocorrência de reuniões para discussão dos resultados obtidos na aplicação da AA na 1ª e na 3ª unidade de 2004 e alteração no padrão de freqüência de realização dessas reuniões em decorrência da AA e no padrão de participação. Esses itens foram escolhidos por indicar, pelo menos, uma tendência à discussão dos diagnósticos obtidos pelo grupo de professores.

Ao final da 1ª unidade, as escolas da AA 2004 foram questionadas sobre a ocorrência de reunião para discussão dos resultados da primeira avaliação externa daquele ano. Independente de quanto tempo a escola já estava envolvida com a AA, a resposta foi maciçamente sim – as equipes haviam se reunido para discutir os resultados (933 escolas entre 1.012, 92%). O mesmo padrão de respostas positivas foi encontrado na 3ª unidade, como pode ser visto na próxima tabela. Como já mencionado, os dados aqui discutidos não são representativos do conjunto de escolas envolvidas pela AA. Entretanto, chamam atenção os percentuais de 92% na 1ª unidade e de 89,7% na 3ª unidade (dentre as 916 escolas com dados válidos) que relataram ter conduzido reunião de discussão de diagnóstico na 3ª unidade de 2004. Essa alta freqüência pode ser explicada porque o grupo respondente é considerado o que buscou cumprir o padrão da AA (aplicando os testes a tempo e encaminhamento o RD para o processamento), mas também é possível estar atrelada ao mecanismo de ritualização na elaboração do RD.

Tabela 53: Ocorrência da reunião entre professores e direção (ou coordenação) na escola para discussão dos diagnósticos feitos após aplicação da AA na 3ª unidade de 2004.

Ocorrência de reunião pós-diagnóstico		**Nº de escolas**	**% válido**
Válidos	Não relata	11	1,2
	Não	83	9,1
	Sim	822	89,7
	Total	916	100,0
Missing		1.083	
	Total	1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade – AD 2004

Idealmente, a discussão em conjunto dos diagnósticos obtidos levaria a equipe escolar a um esforço coletivo no sentido de sanar os problemas. Quase 90% das escolas informam ter cumprido essa etapa, como disposto no padrão de aplicação da AA. Entretanto, para as duas unidades, ter ou não conduzido a reunião de discussão não foi um comportamento associado ao tempo de envolvimento do município com a AA (Gamma, associações muito fracas e não significativas). Também não houve diferença significativa entre as médias, em Português e Matemática, do grupo de escolas que conduziram a reunião para discussão dos resultados e o grupo daquelas que não o fizeram (ANOVA).

Na 3ª unidade, quando perguntada sobre o que ocorreu com a freqüência de reuniões para discussão dos diagnósticos dos alunos, a partir do envolvimento da escola com a AA, a maioria (67,4%) das escolas respondeu que não houve alteração e quase 30% relataram aumento de freqüência, como pode ser visto na próxima tabela.

Tabela 54: Alteração na freqüência de reuniões para discussão de diagnóstico dos alunos em decorrência do trabalho com a AA

Mudança na freqüência de reuniões		**Nº de escolas**	**% válido**
Válidos	Não relata	22	2,4
	Diminuiu	4	,4
	Não foi alterada	618	67,4
	Aumentou	273	29,8
	Total	917	100,0
Missing		1.082	
	Total	1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade

Não houve diferença no comportamento relatado pelas escolas quando considerado o ano de envolvimento com a AA. Em relação ao aumento de participação da escola na busca por soluções para os problemas identificados, 93,7% das escolas responderam que passaram a ser mais participativas como conseqüência do trabalho com a AA, como pode ser visto na tabela a seguir.

Tabela 55: Alteração na participação da escola na busca por soluções para os problemas encontrados.

Mudança na freqüência de reuniões		**Nº de escolas**	**% válido**
Válidos	Não relata	15	1,6
	Não passou a ser mais participativa	43	4,7
	Passou a ser mais participativa	859	93,7
	Total	917	100,0
Missing		1.082	
	Total	1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade

Mais uma vez, a resposta da maior parte das escolas foi positiva em termos do efeito pretendido para a política de avaliação da aprendizagem. Mais uma vez também, é importante olhar esse dado com cautela, visto que o mesmo pode ter sido decorrente de ritualização. Foi feita uma análise do comportamento manifesto das escolas e o ano de envolvimento dos seus municípios com a AA, mas a associação foi fraca (Gamma = 0,248), ainda que significativa (α = 0,024). Dentre os elementos de qualidade da avaliação voltada para uso, a participação da equipe escolar na busca pela solução dos problemas foi a única que, embora fraca, foi significativa.

Ainda sobre o compartilhamento do entendimento da realidade escolar como um uso conceitual da AA, há dois outros elementos: o relato da escola sobre a contribuição da AA para que a equipe refletisse sobre as dificuldades dos alunos e sobre a contribuição no sentido de atrelar essas dificuldades aos planos e práticas implementadas. Idealmente, ao refletir sobre essas questões, as escolas poderiam alterar suas práticas e impactar positivamente o desempenho de seus alunos.

Para resposta a essas questões, fechadas, foram oferecidas quatro alternativas: nada, pouco, suficiente e muito. Diferente das questões anteriores, criadas em observância ao que já tinha sido posto espontaneamente em RD, nessas duas questões a equipe central da AA procurou levantar a posição das escolas em termos do delineado pela política. As próximas tabelas apresentam as respostas das 917 escolas.

Tabela 56: Percepção, por parte da equipe escolar (de 1ª a 4ª série), sobre a contribuição da Avaliação de Aprendizagem na reflexão sobre as dificuldades de seus alunos, a partir do diagnóstico obtido nas unidades. 2004, 3ª unidade.

Contribuição da AA para a reflexão sobre as dificuldades dos alunos		**Nº de escolas**	**% válido**
Válidos	Não relata	20	2,2
	Nada	4	,4
	Pouco	44	4,8
	Suficiente	376	41,0
	Muito	473	51,6
	Total	917	100,0
Missing		1.082	
	Total	1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade

Como pode ser visto na tabela anterior, mais de 90% das escolas relataram uma percepção de efeito positivo (muito e suficiente) da AA para a reflexão sobre as dificuldades dos alunos. Nesse aspecto, tal posição supera muito a predição de uso a partir da categoria Utilidade, levando-se em consideração que esse grupo é de *stakeholder*/usuário foco. Após reorganização das respostas em dois grupos, um negativo (nada e pouco) e outro com percepção positiva (suficiente e muito) e de tratar o não relato como *missing*, foi feito uma comparação das médias *theta* - TRI em Português e Matemática. A diferença encontrada foi muito pequena, como pode ser visto na tabelas a seguir.

Tabela 57: Desempenho em Língua Portuguesa (*theta* TRI) e em Matemática na AD 2004 quando observadas as escolas que consideraram que a AA contribuiu para a reflexão sobre as dificuldades dos seus alunos e aquelas que não tiveram essa percepção.

Local	Percepção das escolas sobre a contribuição da AA para reflexão sobre as dificuldades dos seus alunos	No escolas	Média	Desvio Padrão	Erro Padrão
Língua Portuguesa	Escolas que percebem pouco ou nenhuma contribuição	48	-,6963	,70814	,10221
	Escolas que percebem muito ou suficiente a contribuição	849	-,5517	,75932	,02606
	Total	897	-,5594	,75700	,02528
Matemática	Escolas que percebem pouco ou nenhuma contribuição	48	-,4108	,81931	,11826
	Escolas que percebem muito ou suficiente a contribuição	849	-,1878	,76963	,02641
	Total	897	-,1997	,77352	,02583

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade e AD 2004

Tabela 58: Resultado ANOVA – reflexão sobre as dificuldades dos alunos x média em Língua Portuguesa na AD 2004

ANOVA Reflexão sobre a dificuldade dos alunos - Português	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	,713	1	,713	1,249	,264
Entre grupos	511,122	896	,570		
Total	511,834	897			

Tabela 59: Resultado ANOVA – reflexão sobre as dificuldades dos alunos x média em matemática na AD 2004

ANOVA Reflexão sobre as dificuldades dos alunos - Matemática	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	1,029	1	1,029	1,713	,191
Entre grupos	538,066	896	,601		
Total	539,095	897			

Também a contribuição da AA para o relacionamento dos resultados dos alunos às práticas adotadas pelos professores (análise do mérito realizada pela própria escola) foi percebida positivamente. Nesse caso, 90,2% das respostas válidas indicaram que houve um efeito suficiente ou muito, como pode ser visto na tabela a seguir.

Tabela 60: Percepção, por parte da equipe escolar (de 1ª a 4ª série), sobre a contribuição da Avaliação de Aprendizagem para que relacionassem os resultados alcançados pelos alunos com seus planos de aula e com a sua prática. AA 2004 – 3ª unidade.

Contribuição da AA para o estabelecimento de relações entre o diagnóstico e a prática		Nº de escolas	% válidos
Válidos	Não relata	19	2,1
	Nada	4	,4
	Pouco	67	7,3
	Suficiente	427	46,6
	Muito	400	43,6
	Total	917	100,0
Missing		1.082	
	Total	1.999	

Legenda: F = freqüência e %V = percentual válido
Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade

Após reorganização das respostas em dois grupos, um negativo (nada e pouco) e outro com percepção positiva (suficiente e muito) e de tratar o não relato como *missing*, foi feito uma comparação das médias *theta* - TRI em Português e Matemática. A diferença encontrada foi muito pequena, como pode ser visto nas tabelas a seguir.

Tabela 61: Desempenho em Língua Portuguesa (*theta* TRI) e em Matemática na AD 2004 quando observadas as escolas que consideraram que a AA contribuiu para a reflexão sobre a relação entre suas práticas e planos e o diagnóstico dos seus alunos e aquelas que não tiveram essa percepção.

Local	Percepção das escolas sobre a contribuição da AA para reflexão sobre suas práticas e o diagnóstico dos seus alunos	No escolas	Média	Desvio Padrão	Erro Padrão
Língua Portuguesa	Escolas que percebem pouco ou nenhuma contribuição	71	-,6538	,60702	,07204
	Escolas que percebem muito ou suficiente a contribuição	827	-,5494	,76653	,02665
	Total	898	-,5577	,75539	,02521
Matemática	Escolas que percebem pouco ou nenhuma contribuição	71	-,3138	,58329	,06922
	Escolas que percebem muito ou suficiente a contribuição	827	-,1883	,78904	,02744
	Total	898	-,1983	,77524	,02587

Fonte: Projeto de Avaliação Externa / Base de dados AA 2004 3ª unidade e AD 2004

Tabela 62: Resultado ANOVA – reflexão sobre os diagnósticos dos alunos e os planos e prática utilizados na escola x média em Língua Portuguesa na AD 2004

ANOVA Reflexão sobre dificuldade e prática - Português	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	,950	1	,950	1,659	,198
Entre grupos	512,502	895	,573		
Total	513,452	896			

Tabela 63: Resultado ANOVA – reflexão sobre os diagnósticos dos alunos e os planos e prática utilizados na escola x média em matemática na AD 2004

ANOVA Reflexão sobre dificuldade e prática - Matemática	Soma dos quadrados	df	Mean Square	F	Sig
Dentro dos grupos	2,260	1	2,260	3,788	,052
Entre grupos	533,848	895	,596		
Total	536,108	896			

A partilha é a última dimensão a ser analisada na categoria Uso Conceitual. O panorama oferecido diz respeito a um recorte do público atingido, não só por não abranger todos os *stakeholders*, mas principalmente porque apenas traz o relato daquelas escolas que reagiram à AA a tempo de terem processados seus RD. Ainda assim, as duas subseções Uso Instrumental e Uso Conceitual dão uma noção dos usos de resultados e de outros elementos da AA, complementadas pelas percepções das escolas sobre a política. Com isso, encerra-se a apresentação das contribuições da AA no seu primeiro ciclo. A próxima seção sintetiza os achados nas considerações finais.

5. Considerações Finais

5.1 Uma síntese da pesquisa

Políticas de avaliação são justificadas por suas contribuições ao objeto avaliado. Elas são caras e, na definição da agenda política, competem com outras propostas pelos recursos escassos dos governos. Não há sentido em se implementar uma política de avaliação em larga escala se seus *stakeholders* e usuários não fazem uso de seus resultados ou de outros elementos que possam vir a favorecer o melhoramento do objeto. O presente trabalho, desenvolvido para o doutoramento de seu autor, buscou levantar as contribuições de uma política de avaliação, a Avaliação da Aprendizagem (AA), para seus *stakeholders* / usuários principais: as escolas públicas baianas (consideradas o nível micro de implementação). Foi feita opção por um delineamento metodológico que levasse em consideração os relatos feitos pelas escolas e as taxas (aprovação, reprovação, abandono, desempenho em Português e em Matemática) que as mesmas apresentassem. A pesquisa foi desenvolvida a partir de dados secundários obtidos do Projeto de Avaliação Externa e da Secretaria da Educação do Estado da Bahia. Para analisar algumas das relações mais importantes, foram usados Gamma (quando as variáveis eram ordinais) e ANOVA (na comparação de médias entre grupos).

Considerando que o foco da investigação foi dirigido para as contribuições das políticas de avaliação, é importante registrar que, apesar de usar uma categoria da meta-avaliação para a análise dessas contribuições, a pesquisa em relato não deve ser classificada como meta-avaliação. Não houve a pretensão da determinação de um padrão para o julgamento do objeto ou sua para aplicação. O foco do estudo também não foi voltado para a discussão sobre a capacidade de predição de uso dos itens de verificação na categoria Utilidade. Registrados esses limites, o presente capítulo apresenta considerações finais que apontam certas tendências de contribuições para o objeto avaliado. Finalizando o relato da pesquisa, o objetivo dessa seção é, além de sintetizar o texto anterior, apresentar reflexões sobre os achados relacionados na Seção 4.

5.2 As contribuições da Avaliação de Aprendizagem

A política escolhida como foco da presente investigação foi a vertente Avaliação de Aprendizagem (AA) do Projeto de Avaliação Externa, inserido no programa de reforma do Governo da Bahia nos anos de 1999 a 2004 como ação prioritária (ambos descritos na Subseção

3.1). O estudo focalizou as contribuições do primeiro ciclo da AA concretizado no período de 2001 a 2004. Justificou-se a escolha da AA por quatro razões:

1) Concordou-se com Souza (2002: s/p) na assunção de que programas e projetos do governo são políticas públicas "postas em prática". Dada essa conceituação a AA, implementada em larga escala sob a coordenação do ISP na UFBA, foi analisada como uma política pública. A escolha do estudo de suas contribuições no nível micro de implementação – a escola – deveu-se ao entendimento de que o delineamento final de qualquer política é dado não pelos formuladores, mas pelos seus implementadores, especialmente aqueles em contato direto com o cidadão (LIPSKY, 2000). Quanto a esse aspecto, assumiu-se que os desvios dos usos pretendidos ou usos feitos por usuários não previstos originalmente na formulação da política não são ruins *per se*, não sendo necessariamente fruto de má conduta, e que poderiam enriquecer a explicação sobre as contribuições das políticas de avaliação.

2) A AA foi especialmente interessante porque, enquanto política, afastou-se do formato mais freqüente utilizado pelas avaliações implementadas pelo Governo Federal e por vários governos estaduais e municipais no contexto da reforma do estado. Em lugar de concentrar-se nas questões sobre *accountability* ou transparência pública, que justificavam os programas de avaliação à época, a AA empregou seus esforços para o favorecimento de diagnósticos detalhados em duas disciplinas básicas. Fez isso a partir do pressuposto que, em um panorama de autonomia escolar fortalecida, a equipe docente teria condições para replanejar seus cursos e melhor atender o alunado. No caráter diagnóstico da AA residiu o aspecto do valor da política. O estudo sobre ela poderia subsidiar novas políticas.

3) A AA teve finalizado o seu primeiro (e único) ciclo em 2004. Naquele ano, a Avaliação de Desempenho (AD) seria aplicada a alunos de 4ª série que teriam participado da AA desde 2001, caso não tivessem sido reprovados ou não tivessem abandonado o curso. O ciclo da AA poderia, então, ser analisado a partir de indicadores externos, como a variação nas taxas de aprovação, reprovação e abandono das escolas no período sob investigação (2001 a 2004), como também pelo desempenho dos alunos em Português e em Matemática.

4) O delineamento da AA permitiu o registro de respostas das escolas envolvidas com a política para perguntas que lhes foram encaminhadas a cada aplicação de provas desde o seu início, em 2001. Essas respostas foram categorizadas e tabuladas pela equipe central da AA a partir de 2002 e favoreceriam a análise das suas contribuições

> para além do uso dos resultados. A hipótese levantada na pesquisa em relato foi: “em políticas de avaliação educacional em larga escala, os resultados são elementos pouco utilizados e é o acontecimento da avaliação que afeta as instituições em nível micro (escolas)”. Os dados da AA, já tratados e sistematizados, possibilitaram, respeitados os limites de tempo e custo, a verificação da hipótese.

No contexto geral de reforma educacional proposto pelo Governo da Bahia em 1999, a AA, inserida em um programa maior, deveria colaborar para a melhoria da qualidade da educação (especialmente a pública) do Estado. Mais especificamente, ao fornecer instrumentos de avaliação a cada 200 horas letivas, a AA favoreceria o diagnóstico de problemas em cada turma do Fundamental Menor a tempo de saná-los. Idealmente nesse processo, os alunos – mais preparados – teriam maior aprendizagem (expressa em desempenhos em Português e Matemática), perderiam menos o ano e deixariam de abandonar suas escolas.

As contribuições da AA foram sistematizadas em duas categorias: Utilidade e Uso. Após análise da AA por meio de seus documentos oficiais de divulgação e de seus relatórios técnicos, buscou-se levantar os relatos das escolas sobre a percepção de utilidade e sobre os usos concretizados durante o ciclo. Também foi examinado se tais usos guardavam alguma relação com o tempo de envolvimento na política ou com os resultados dos alunos em Português e em Matemática. Os pressupostos foram: as escolas que tivessem cumprido um ciclo completo (ou quase) da AA já teriam visto seus efeitos sobre as quatro séries avaliadas (1ª a 4ª séries do Fundamental Menor) e saberiam usar melhor a ferramenta da avaliação. Alguns comportamentos (como planejar o curso a partir dos diagnósticos do ano anterior, discutir os resultados em grupo de docentes, ou reforçar, durante o ano letivo, as estratégias de ensino para lidar com os problemas identificados pela AA) teriam sido internalizados em maior freqüência nessas escolas. Isso posto, os resultados dos seus alunos em Português e Matemática na vertente de Avaliação de Desempenho seriam mais altos que aqueles de alunos de escolas que não haviam tido envolvimento com a AA. Nessa mesma lógica, alunos melhor preparados passariam de ano, o que influiria para a queda das taxas de reprovação e de abandono e para o aumento das taxas de aprovação nas redes estadual e municipais.

Ao todo, foram utilizados dez indicadores (U1 a U8 e Usos Instrumental e Conceitual), observados em dezoito dimensões, ainda que com abordagens diversas. Os indicadores U1 a U7, adaptados do *checklist* proposto por Stufflebeam para a meta-avaliação a partir de padrões do JCSEE, foram analisados com base na identificação, na AA, da presença de 66 itens verificadores.

O comportamento da AA para o Indicador U8 foi levantado por meio de relatos das escolas. O mesmo se deu para análise da AA nas dimensões de Uso Conceitual. Já as contribuições da AA nas dimensões de Uso Instrumental valeram-se de taxas de eficiência (aprovação, reprovação e abandono) da 1ª e da 4ª séries em 2001 e 2004 e de desempenho (média *theta*-TRI, por escola) dos alunos de 4ª série em Português e em Matemática na AD 2004. A ilustração a seguir sintetiza as categorias e seus indicadores.

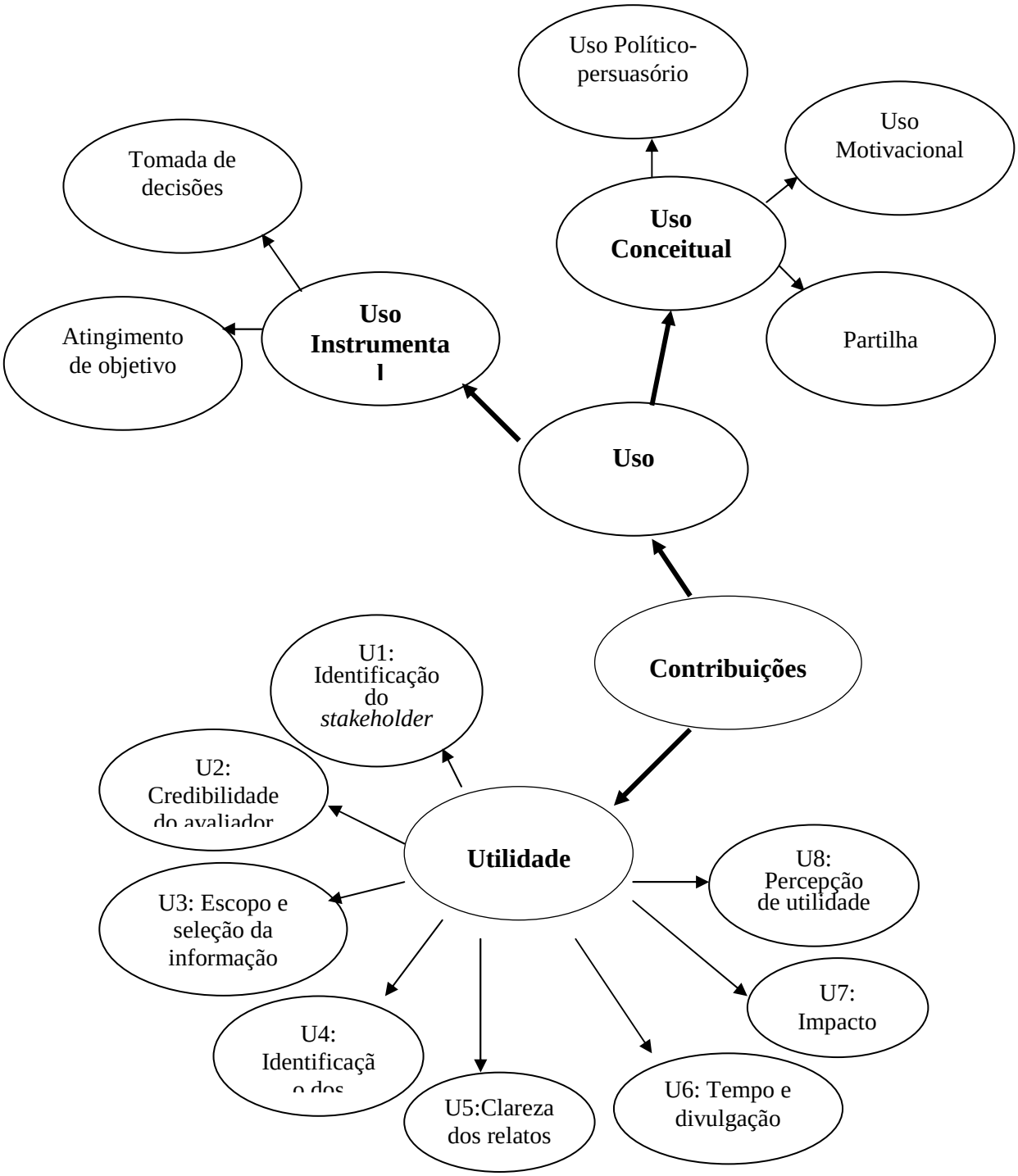


Ilustração 21: Representação dos oito indicadores da categoria Utilidade e dos dois indicadores da categoria Uso utilizados para análise das contribuições da política de Avaliação de Aprendizagem em 2001 – 2004.

Os resultados foram apresentados na Seção 4. Especialmente os relatos das escolas foram restritos àquelas unidades que encaminharam, a tempo de processamento, os Relatórios do Diretor (RD) para a equipe central da avaliação. Esse percentual girou em torno de 50%. Os dados sobre desempenho dos alunos, oriundos da Avaliação de Desempenho, incluíram as escolas urbanas que haviam preenchido os RD, bem como as demais. Para análise das taxas de aprovação, reprovação e abandono, no entanto, a base Censo SEC/MEC favoreceu uma análise contrastiva entre todas as escolas, organizadas por sua localização: rural ou urbana. A seguir são sintetizados os resultados mais expressivos.

1) Dentre os 66 itens verificadores dos sete indicadores da categoria Utilidade (U1 a U7), 44 foram observados na AA (67%). No *checklist* original de Stufflebeam, tais itens apontam a qualidade do delineamento avaliativo por facilitarem seu uso. Foi assumido o pressuposto de que quanto mais itens da Categoria Utilidade tivessem sido observados na política de avaliação, tanto maior seria sua probabilidade de uso. No caso em relato, considerou-se que a probabilidade seria média, tendendo para alta.
2) No presente estudo, à categoria Utilidade foi acrescido um oitavo indicador: a percepção da utilidade que os *stakeholders* constroem sobre a experiência avaliativa, que independe das finalidades da avaliação e dos usos concretizados. Especialmente nas respostas ao RD em 2004, a AA foi percebida como útil pelo grupo respondente (75% escolas dentre as 42% delas que encaminharam os RD). Se a AA foi percebida como útil, de alguma maneira foi utilizada ou havia uma expectativa alta para seu uso.
3) Da análise da AA nos oito primeiros indicadores (categoria Utilidade), descrita na Subseção 5.1, alguns aspectos devem ser ressaltados:
    a) Dentre os *stakeholders* e usuários da AA identificados no U1, foram atendidos o cliente e a escola, *stakeholder* principal no delineamento da política. Os demais *stakeholders* foram menos atendidos. Em especial, a decisão de implementação da AA apenas em escolas urbanas excluiu um grande espectro de escolas rurais, muito importantes para a formulação de políticas pela administração municipal.
    b) A decisão de sistematização de dados (não só da AA) para utilização de usuários indiretos, como a Academia, resultou em trabalhos de iniciação científica, mestrado e doutorado. Apesar de indireta, essa foi uma contribuição concreta da avaliação.
    c) As escolas não participaram do processo de formulação da AA, mas tiveram centralidade em sua implementação. Vieram das escolas avaliadas os professores que elaboraram as matrizes de referência e os itens de testes; a administração e a

elaboração dos diagnósticos estiveram sob a responsabilidade de cada unidade escolar; especialmente, a tomada de decisões a partir dos resultados da AA esteve concentrada no nível micro. Entretanto, o retorno do RD à equipe central da avaliação não ocorreu plenamente, sendo essa a etapa de finalização de cada aplicação. Uma reflexão merece ser feita: a literatura recente defende que uma avaliação participativa seja uso-conducente. Na análise da experiência em tela, essa relação não é evidente. Há diversos fatores que interferem no uso da avaliação, participativa ou não, sendo um deles a capacidade institucional. Como já conceituado, tal capacidade é a "habilidade de compreender e analisar uma determinada situação, identificar problemas, definir e implementar metas, objetivos e formular estratégias para ações futuras" (CALMON, 2005:6). Com alguma freqüência percebeu-se que as escolas que encaminharam seus RD foram capazes de identificar os problemas e até de definir ações futuras, mas sem uma relação com o diagnóstico feito. Também foram encontrados pedidos freqüentes de ajuda, encaminhados pelas escolas à equipe central, fora do escopo da política de avaliação (como capacitações, soluções para problemas com docentes e discentes ou infra-estrutura, materiais didáticos, dentre muitos). Por outro lado, a burocracia média no órgão central da Educação do Estado (SEC) não acreditava ser a escola capaz de seguir sem tutoria, sendo esse o principal argumento (manifesto) contra o fortalecimento da autonomia. No entanto, mesmo com as representações regionais, o Estado não tinha capilaridade suficiente para realizar a tutoria defendida. Sem capacidade instalada, não há uso da avaliação; contudo, a centralização da capacidade em um único local (SEC-Salvador em um Estado com 417 municípios) também não leva ao uso. As políticas de fortalecimento de autonomia (no contexto de descentralização) – para as quais a avaliação é dada como contrapartida – deveriam considerar a instalação da capacidade local. Um delineamento avaliativo como o utilizado pela AA demanda um bom repertório para mudança no nível da tomada de decisões, sob pena de não contribuir para o atingimento da finalidade da política. Por essa razão, sugeriu-se aqui a criação de um novo item verificador para o Indicador U7 (Impacto): "Identifica o repertório para mudança dos *stakeholders*".

d) A falta de articulação com as demais políticas do órgão central (SEC) e a dificuldade de articulação do Estado com os municípios, especialmente em relação às questões pedagógicas, de alguma maneira contribuíram para que a AA tivesse sido implementada de maneira isolada. Esse isolamento pode ter sido acentuado

visto que não se perceberam relatos, por parte da escola, do uso (conceitual) dos diagnósticos para pressionar os órgãos centrais (estadual ou municipais) por apoio às propostas de saneamento dos problemas identificados. Além disso, a queda do percentual de encaminhamento dos RD no quarto ano pode indicar um cansaço da escola que, por um lado, não tinha repertório para sanar seus problemas e, por outro lado, não recebeu apoio externo para fazê-lo. Nesses casos, cumprir o padrão da AA tornou-se sem sentido, a não ser na ritualização dos procedimentos avaliativos como estratégia de sobrevivência às determinações do órgão central (LIPSKY, 1980; DANTAS, 2005). Na presente pesquisa, defende-se essa ritualização como um "não uso", a partir da distinção feita por Patton (1988b *apud* SHULHA) entre uso x não uso e mau uso x não mau uso. Da leitura dos relatos das escolas, percebeu-se a ritualização como um movimento voltado às estratégias para lidar com as demandas da AA (não uso), bastante diverso de uma manipulação de dados ou de processos para, por exemplo, maquiagem dos resultados (mau uso, atrelado à má fé).

e) Houve uma demanda não atendida, especialmente no nível municipal, de avaliação com fins de controle e regulação. Esse delineamento seria a contrapartida para uma escolha dos dirigentes escolares feita pelo prefeito e não pela secretaria municipal da educação. Mais adiante, foi interessante perceber, como uso conceitual relatado, o uso da AA para o monitoramento do trabalho docente, não previsto e não recomendado originalmente no formato da política. Interessantemente, o discurso predominante nas secretarias de educação e escola tem natureza inclusiva e participativa, oposto à demanda por avaliação para controle.

f) Um delineamento de caráter *low stakes*, como o utilizado pela AA, foi considerado ideal para implementação da cultura de avaliação, já que, ao tempo que fornecia uma ferramenta diagnóstica às escolas, não expunha publicamente seus resultados. O monitoramento de uma amostra na aplicação dos testes e a análise dos RD apontam para desvios de implementação da política que, quando existentes, raramente foram associados à má fé. No conjunto, foram adequações das aplicações à realidade escolar ou ainda resposta a problemas logísticos (como atrasos ou falta/sobra de materiais). A decisão por um caráter *low stakes*, no entanto, provavelmente teve como uma das conseqüências o percentual médio (média/ano: 49% em 2001, 57% em 2002, 58% em 2003 e 48% em 2004) de encaminhamento dos RD após a aplicação das provas. As análises dos resultados

da AA demandavam, das equipes escolares, um grande trabalho a cada 200 horas letivas que, sobreposto à carga já existente, pode ter sido um elemento de desmotivação. Visto que os resultados eram internos, é possível que muitas escolas tenham desistido a) de aplicar os testes e/ou b) de elaborar os RD. Merece reflexão, em delineamentos avaliativos futuros, a decisão sobre o caráter *low* ou *high stakes* de uma determinada avaliação. O fato de ser *low stakes* favorece a implementação da política a curto prazo; entretanto, pode dar a impressão de que os resultados obtidos, por não serem acompanhados por "ninguém", também não "levam a nada". Isso é especialmente importante nas políticas de avaliação que são conduzidas de maneira isolada, não articuladas com outras políticas que busquem o melhoramento. Por outro lado, avaliações *high stakes*, como discutido no marco teórico, pressupõem um monitoramento bem mais pesado da coleta de dados, para evitar que desvios alterem os resultados. Além disso, avaliações *high stakes* isoladas, sem articulação com outras políticas que ajudem o *evaluand* a melhorar o objeto, podem ser consideradas cruéis e excludentes.

g) O estabelecimento do padrão para o julgamento de valor foi feito a partir de uma abordagem estatística (Angoff modificado) para um desempenho das turmas nas dimensões de conteúdo informadas por matrizes de referência. A AA, por um lado, contribuiu para a disseminação desses conteúdos; por outro, viu surgir uma resistência a eles por "não refletirem a realidade das escolas públicas" ou do "aluno pobre". Por mais que as linhas de corte (estabelecidas por teste) representassem uma proficiência na unidade letiva, várias escolas as perceberam negativamente.

h) Dois dos sete Indicadores da categoria Utilidade original do JCSEE foram dedicados a aspectos da comunicação. Não há disputa no fato de que uma avaliação só será útil se seus usuários tiverem acesso a ela. A AA teve um comportamento especialmente interessante no U5 (Clareza no relato da Avaliação) e um pouco menos positivo no U6 (Tempo e divulgação dos relatórios) quando o *stakeholder*/usuário considerado foi a escola. Problemas decorrentes da perecibilidade da informação para a tomada de decisão foram sanados à medida que a própria escola pode produzir seus resultados. Duas reflexões merecem ser feitas em termos das contribuições da AA: 1) os tempos de entrega dos relatórios aos demais *stakeholders* podem ter cooperado para o isolamento da escola na solução dos seus problemas. As secretarias municipais da Educação só passaram a receber as sínteses de monitoramento em 2004. A SEC os recebeu desde sempre,

mas na forma de relatórios técnicos especializados. Para próximos delineamentos, é importante estabelecer um sistema no qual não só os usuários principais tenham acesso à informação no tempo adequado, mas também outros *stakeholders* que possam apoiá-los. Para delineamentos *low stakes*, os dados para outros *stakeholders* podem ser agregados, sem implicação direta por escola, mas com mapeamento de regiões, área (urbana x rural), etc. O segundo ponto diz respeito ao meio adequado de comunicação com a escola pública. Dada sua heterogeneidade (em termos de linguagem, de infra-estrutura) e aos custos da comunicação para um público tão abrangente quanto distribuído espacialmente, os contatos por meio impresso são hoje os mais acessíveis. É interessante, no entanto, considerar a *internet* como o meio a ser utilizado (desde, obviamente, que as escolas sejam ligadas em rede).

4) O padrão de respostas das escolas tanto em relação à percepção da utilidade da AA quanto do uso dos seus elementos pode indicar comportamento ritualizado, já discutido. Nesse sentido, poderia ser considerado um não-uso. Uma das suspeitas de comportamento ritualizado foi relacionada ao encaminhamento adequado do RD para a equipe central. Interessantemente, quando comparados os desempenhos médios dos alunos de 4ª série em Língua Portuguesa e em Matemática na AD 2004 pelos dois grupos de escolas (as que encaminharam o RD a tempo na 3ª unidade de 2004 e as que não o fizeram), foram encontradas diferenças significativas. Embora, de forma geral, o desempenho tivesse sido sofrível, as escolas que encaminharam os RD tiveram médias acima daquelas que não o fizeram. Seja porque os alunos aprenderam a fazer as provas (e não necessariamente a dominar o conteúdo), porque as escolas se adequaram ao formato AA, ou ainda porque a AA realmente contribuiu para que as dificuldades diagnosticadas fossem sanadas, os resultados da AD2004 foram mais satisfatórios no grupo que cumpriu as etapas previstas pela AA.
5) Quando perguntadas em 2004 se haviam utilizado a AA para planejar o curso em 2004 (uso instrumental pretendido), o grupo respondente, em sua maioria, afirmou que sim, tanto os resultados (56%) quanto (e principalmente) as matrizes de referência (66% 1ª e 2ª série e 71% 3ª e 4ª série). Esse uso relatado pelas escolas contrariou a expectativa geral, vista na literatura, de não utilização de dados em uma avaliação em larga escala. Mais da metade do grupo respondente indicou a forma clássica de utilização da avaliação (uso de resultados para a tomada de decisão). Entretanto, o padrão de respostas das escolas apontou também no sentido da confirmação da hipótese de pesquisa de que “em políticas de avaliação educacional em larga escala, os resultados são elementos pouco utilizados e é

o acontecimento da avaliação que afeta as instituições em nível micro (escolas)". As matrizes de referência foram mais usadas na tomada de decisão que os próprios diagnósticos. Se, por um lado, essa contribuição da AA foi a esperada, por outro, alguns problemas podem ter sido a ela associados: i) ritualização; ii) ensino para o teste; iii) falta de capacidade instalada. A utilização das matrizes pode ter estado relacionada a uma redução do currículo escolar àquilo que "cai no teste", bem como à incapacidade da escola de propor seu próprio currículo. Essas questões merecem uma investigação mais aprofundada.

6) Ainda em termos da utilização da AA na sua forma instrumental clássica, um dado é curioso: não foi possível estabelecer uma relação entre o uso da AA no planejamento e o desempenho dos alunos. Esse achado suscitou a seguinte posição: o uso instrumental – tão defendido como aquele que justifica a avaliação – não necessariamente resulta no atingimento da finalidade de uma política.
7) O atingimento da finalidade da política foi considerado como uso instrumental quando, em verdade, deveria ser tratado como uma conseqüência dos diversos usos em uma avaliação. Visto que o presente trabalho não buscou relações de causa x efeito, optou-se por buscar tendências nas taxas de aprovação, reprovação, abandono e desempenho que pudessem ser, de alguma maneira, indicativas da relação das escolas com a AA. Por essa razão, o atingimento foi considerado um uso.
8) A análise da variação das taxas de aprovação, reprovação e abandono indicou que, de maneira geral para as escolas públicas rurais e urbanas na Bahia, o comportamento foi oposto ao esperado pelo menos quanto às duas primeiras. As taxas de aprovação foram menores em 2004 que em 2001 e as taxas de reprovação foram maiores em 2004 que em 2001. Esses comportamentos, entretanto, foram comuns às escolas que participaram e que não participaram da AA. Vale ressaltar que a queda na taxa de aprovação não pode ser entendida – *a priori* – como algo ruim. Em 2001, havia uma percepção de aprovação em massa de alunos que não detinham as competências mínimas para acompanhar as disciplinas nas séries seguintes. Uma hipótese de explicação dessas variações pode estar relacionada a uma postura mais crítica das escolas desenvolvida nos quatro anos do programa de reforma de Estado.
9) O aspecto mais positivo na observação das taxas foi a queda na taxa de abandono da 1ª e da 4ª séries. Os alunos, apesar de uma maior reprovação, abandonaram menos a escola. Mais uma vez, entretanto, não se pode afirmar que essa foi uma contribuição da AA ou mesmo do programa de reforma.

10) Quando comparados os desempenhos em Língua Portuguesa e em Matemática dos alunos de 4ª série na AD 2004, os resultados daquelas escolas situadas em municípios envolvidos pela AA na totalidade do ciclo ou quase foram superiores àqueles oriundos de escolas cujo envolvimento se deu apenas em 2004. No caso da 4ª série em Matemática, a diferença foi maior que aquela encontrada em Português.
11) Dos itens 4 e 10 acima poder-se-ia pensar na associação AA – maior aprendizagem em Português e em Matemática, refletida em maiores médias de desempenho nas provas da AD 2004. Caso essa associação seja real, é possível considerar-se que a política atingiu seu principal objetivo – melhorar a qualidade do ensino público -, especialmente nos aspectos do desempenho dos alunos. O valor da AA (considerando que teria tido mérito), partindo de sua característica diagnóstica, poderia ser assim associado à contribuição – social – de uma educação de maior qualidade. Entretanto, essa leitura deve ser feita com muito cuidado. As próprias escolas informaram perceber que os alunos se acostumaram ao formato dos testes ao longo dos anos. Esse fator pode ter tido efeito sobre sua capacidade de resposta a questões de múltipla escolha, que é diferente de uma maior competência nas duas disciplinas.
12) Em termos do uso conceitual, não foi possível estabelecer qualquer relação forte ou significativa entre os relatos das escolas e o tempo de envolvimento de seus municípios com a AA ou entre esses relatos e o desempenho dos alunos. Apesar de, para vários itens verificadores, os resultados terem indicado contribuições da AA, novamente é necessário cautela na análise dos mesmos e, mais uma vez, essa cautela é decorrente de o grupo respondente não ser representativo do todo e da sombra de um comportamento ritualizado que paira sobre os RD respondidos.
13) Dentre os efeitos não instrumentais da AA, um foi relacionado à formação (ou fortalecimento) de um traço crítico na escola que a faria capaz de perceber seus pontos fracos. A identificação de necessidade de capacitação docente nas disciplinas sob avaliação ocorreu em boa parte das escolas, mas disso não resultou uma pressão por apoio dos órgãos centrais ou pode ser associado ao desempenho do alunado. As escolas informaram uma demanda de capacitação em Português superior àquela de Matemática, ainda que essa última tenha sido a disciplina com o desempenho mais insatisfatório. Além disso, o traço crítico que permitiu à escola perceber que necessitava de capacitação pode estar relacionado ao mesmo movimento que resultou em taxas de aprovação mais baixas em 2004, não sendo, portanto, associado à AA.
14) Um aspecto intrigante identificado nos relatos das escolas disse respeito à avaliação como elemento de motivação para a aprendizagem. A literatura sobre avaliação discute

enormemente o quão excludente e injusta ela pode ser, mas, no caso em relato, as escolas tiveram a percepção de que seus alunos estavam mais motivados a aprender em conseqüência da AA. Observando-se esses relatos por série tem-se que, na 4ª série, quase 80% dos respondentes consideraram esse efeito. Como nos outros casos, não foi possível estabelecer qualquer relação entre esse registro e o tempo de envolvimento com a AA. Também não houve uma relação entre tais relatos e o desempenho dos alunos nas disciplinas sob testagem pela AD.

15) O uso da AA para o compartilhamento de uma visão de realidade foi, para o Indicador Uso Conceitual, o mais relevante. Um percentual muito alto dos respondentes (92% na 1ª unidade e 89% na 3ª, em 2004) afirmou a ocorrência das reuniões para discussão do diagnóstico e mais de 90% das escolas relataram uma percepção de efeito positivo da AA para a reflexão sobre as dificuldades dos alunos. Também a contribuição da AA para o relacionamento dos resultados dos alunos às práticas adotadas pelos professores (análise do mérito realizada pela própria escola) foi percebida positivamente por 90,2% dos respondentes. Em relação ao aumento de participação da escola na busca por soluções para os problemas identificados, 93,7% das escolas responderam que passaram a ser mais participativas como conseqüência do trabalho com a AA. Dentre os elementos de uso conceitual, a participação da equipe escolar na busca pela solução dos problemas foi a única que, embora fraca (Gamma = 0,248), foi significativa (α = 0,024) em termos do envolvimento do município com a AA. Nesse aspecto, tal posição supera muito a predição de utilidade a partir da categoria Utilidade, levando-se em consideração que esse grupo é de *stakeholder*/usuário foco.

Os quinze pontos acima relacionam as principais contribuições da AA quando observado um ciclo completo de implementação. De maneira geral, pode-se dizer de uma utilização dos elementos da avaliação muito mais ampla que aquela restrita aos resultados. Respeitados os limites da não representatividade dos respondentes e de uma possível ritualização nas respostas das escolas, os achados apontaram para um uso real da avaliação, com efeito no desempenho do alunado da 4ª série tanto em Português quanto em Matemática.

A busca das contribuições de AA para seus *stakeholder*s principais favoreceu uma reflexão sobre o modelo de meta-avaliação proposto por Stufflebeam com base nos padrões de qualidade do JCSEE, adaptado para análise de políticas públicas de avaliação. A subseção a seguir apresenta algumas considerações sobre esse modelo.

5.3 Uma reflexão sobre o modelo utilizado

O JCSEE (1994) definiu 30 padrões de qualidade para a avaliação de programas a partir de 04 categorias: Precisão (12 padrões), Viabilidade (03), Propriedade (08), e Utilidade (07). Stufflebeam (1999) propôs um *checklist* para que fosse possível verificar o atendimento aos padrões, detalhando cada um deles em 10 itens verificadores, em um processo de meta-avaliação. No presente trabalho, foi feita uma adaptação da categoria Utilidade do *checklist* para aplicação em políticas públicas de avaliação, implementadas em larga escala.

Os resultados do presente estudo, relacionados na Subseção 4.1, são apresentados em um texto redundante, no qual um mesmo aspecto é observado várias vezes, ainda que por ângulos diversos. A redundância do texto reflete um problema grande com o modelo utilizado: o fato de os itens verificadores estarem repetidos em vários padrões da categoria Utilidade (e mesmo em outras categorias). Além disso, os padrões na referida categoria não estão todos no mesmo nível analítico. Alguns, como o U5 e o U6, buscam aspectos facilmente associados a questões técnicas da avaliação e poderiam ser alocados em uma categoria que lidasse com o mérito; outros, como o U1 e o U7, pressupõem uma análise que ultrapassa a verificação direta dos itens e estão mais ligados ao valor da política de avaliação.

Quanto ao referido modelo, é importante registrar que o número de itens de verificação é muito grande, além do necessário para apontar a qualidade da experiência. Concorda-se, portanto, com as críticas feitas por Widmer (2005) a partir da aplicação dos padrões a uma experiência suíça. Também é importante relatar que a aplicação não conduz o pesquisador a um nível confortável de precisão, seja pela quantidade de itens sem uma hierarquia de importância para a determinação da qualidade, seja porque, a depender da escolha sobre os *stakeholders* principais, a resposta por item verificador pode variar. Um exemplo para ilustrar o problema: a investigação em relato optou por focalizar a análise da Utilidade da AA para seus *stakeholders*/usuários principais; se tivesse enfatizado os setores técnicos das secretarias municipais, o conjunto de resposta teria sido diferente (a exemplo do U6).

Talvez o uso ideal do *checklist* possa ser atrelado à formulação de políticas de avaliação: nesse caso, a repetição dos itens apenas reforçaria a importância de se levar em consideração esse ou aquele aspecto no seu delineamento. Tais aspectos, mesmo não considerados como preditores, certamente podem ser relacionados como facilitadores do uso. Além disso, a repetição pode ser interessante em situações de formação do avaliador/equipe avaliadora.

Por fim, a categoria não deveria ser nomeada Utilidade, vez que esse conceito está imbricado à percepção do usuário. Da mesma forma, o U7 não deveria ser chamado Impacto vez que trata da relação avaliação x *stakeholder* muito mais que das mudanças observadas no objeto avaliado.

Pensando-se em um modelo com variáveis ideais para a meta-avaliação de políticas de avaliação, as mesmas deveriam ser mutuamente exclusivas, em número reduzido, exaustivas e pertinentes. Dentre elas, deveria estar "considera o repertório para mudança" já discutido anteriormente, além de itens voltados para percepção de utilidade dos *stakeholders* e de usos concretizados. Por fim, é necessária a definição de uma escala com descrição dos níveis, de modo que, ao meta-avaliador seja possível fazer um julgamento de valor sobre valor e mérito da avaliação sob estudo.

Feitas essas reflexões sobre o modelo utilizado, a última subseção sintetiza as contribuições da presente pesquisa.

5.4 As contribuições deste trabalho

A literatura sobre a avaliação é rica em manifestos sobre a baixa utilização das avaliações e, em especial, das avaliações em larga escala. A análise da AA mostra que, para o grupo que encaminhou os RD, houve uso da avaliação, pretendido ou não. Diante do quadro apresentado pela literatura, não deixa de ser surpreendente o relato do uso instrumental dos resultados, bem como de outros elementos da AA. Por essa razão, advoga-se aqui a ampliação da noção de uso instrumental para além dos resultados. A aplicação de cadernos de teste excedentes para exercício em sala de aula ou a incorporação da análise dos resultados para a turma (em lugar do foco no aluno) nas avaliações regulares das escolas são exemplos de usos instrumentais que não estão relacionados à tomada de decisões a partir dos resultados. Mesmo em uma situação de ensino direcionado para o desempenho nos testes, deve ser considerado uso a incorporação das matrizes de referência no currículo das escolas no Ensino Fundamental, quase como um "efeito vestibular" observado no Ensino Médio. A discussão não deveria ser feita sobre uso x não uso; deveria ser concentrada em análises sobre se os usos feitos contribuem ou não para a melhoria da qualidade da Educação.

Por outro lado, como argumentado por Weiss, usos não instrumentais também são fundamentais para o entendimento de como uma política de avaliação pode afetar seus *stakeholders*. Em especial, o estudo da AA mostrou a força da política para o compartilhamento de uma visão sobre a realidade de ensino, sendo este o uso conceitual mais relatado, de maneira positiva, pelas

escolas. Tivesse o modelo de análise sido concentrado apenas no uso instrumental clássico e estes efeitos da política seriam desconhecidos.

A segunda contribuição do presente trabalho diz respeito à reflexão sobre a relação entre uso e o atingimento da finalidade da política de avaliação. Diferente do que possa parecer, tal relação não é direta ou linear. A falta de uma associação entre a tomada de decisões em 2004 com base nos diagnósticos feitos em 2003 e o melhoramento das taxas, no caso da AA, levanta essa questão: o uso concretizado, mesmo o instrumental, não necessariamente leva ao atingimento do objetivo maior ou mesmo dos objetivos específicos da política.

Em um contexto de escassez no qual os governos precisam otimizar os recursos, a avaliação, em papel central, teve seu foco deslocado do processo para o produto. O mesmo deveria ocorrer com a meta-avaliação de políticas de avaliação: interessa menos como se dão os usos; o importante é buscar o atingimento da finalidade da política, especialmente em um panorama de descentralização. Como, no nível micro, o implementador transforma a política originalmente formulada, o formulador poderá sem dúvida fazer propostas de certos usos para determinados usuários, como proposto por Patton (*intended uses by intended users*). Entretanto, se os usuários propõem usos diversos ou se fazem ou deixam de fazer uso da forma pretendida é menos importante que o atingimento da finalidade. Esse argumento fica ainda mais forte quando se admite, como já discutido no Marco Teórico, que a tomada de decisões não é necessariamente um processo racional.

Na etapa de formulação de qualquer política de avaliação, espera-se que sejam considerados os fatores facilitadores de uso, como os relacionados nos padrões do JCSEE. Como posto por Ginsburg e Rhett (2003), os delineamentos não garantem os usos, mas devem ser tais que aumentem a probabilidade de que eles ocorram. Não resta dúvida que aspectos como a comunicação de forma direta e clara ou o respeito aos tempos dos usuários a partir do entendimento dos resultados da avaliação como perecíveis, ou ainda perguntas pertinentes cujas respostas sejam do interesse direto dos *stakeholders* sejam uso-conducentes.

Há, no entanto, outros cuidados tão ou mais importantes para o delineamento de políticas de avaliação. O primeiro deles é a articulação dessas políticas com as demais. Sem esse vínculo, a avaliação fica isolada. Concorda-se com Ravela e outros (2008), para quem falta uma maior articulação entre avaliação, desenvolvimento curricular, formação inicial e desenvolvimento profissional dos docentes. No caso em tela, talvez pela decisão de que os resultados das escolas

não teriam divulgação feita pelo Projeto de Avaliação, as secretarias municipais muito frequentemente não apoiaram as unidades escolares na busca de reversão do quadro diagnosticado a cada aplicação dos testes. O mesmo se deu com a SEC. Sem articulação entre as políticas, volta-se ao cenário vislumbrado por Helene e registrado na Introdução a esse documento: “como um Narciso às avessas, ficaremos a contemplar a feiúra de nosso sistema educacional, sem intervir, até sermos inteiramente consumidos” (HELENE, s/d: 12).

O segundo cuidado diz respeito a um delineamento que considere a capacidade institucional ao nível da tomada de decisão ou, dito de outra forma, ao repertório para a mudança. De nada adianta um processo avaliativo que resulte em boa informação e em base para o julgamento de valor sobre dados relevantes se aqueles que recebem tais informações não detêm competências ou poder para mudar o que precisa ser mudado. Pelo mesmo motivo propõe-se o terceiro e último cuidado: o delineamento avaliativo deve ser tal que informe o usuário dentro do seu raio de autonomia. Exigir do *stakeholder* / usuário que tome decisões para além das suas possibilidades de ação presta-se apenas para aumentar a frustração dos professores e diretores da rede pública.

Com essas contribuições, conclui-se o relato da pesquisa em tela e a apresentação da tese de doutoramento na esperança de que venham a ser úteis para a formulação de novas políticas de avaliação e para o refinamento de políticas públicas existentes.

Referências

ABRAMOWICZ, Mere. Avaliação, tomada de decisões e políticas: subsídios para um repensar. **Estudos em Avaliação Educacional**, n.10. São Paulo: FCC, jul/dez. 1994, p 81-102.

AFONSO, Natércio. A regulação da educação na Europa: do Estado Educador ao controlo social da escola pública. In: BARROSO, João (Org.). **A escola pública: regulação, desregulação, privatização**. Porto: Edições Asa, 2003, cap. 1, p. 49-78.

AMERICAN EVALUATION ASSOCIATION. **Guiding principles for evaluators**. Publicado em 1994 e revisado em 2004. Disponível em www.eval.org. Acesso em fevereiro de 2008.

ANDRIOLI, Antonio Inácio. As políticas educacionais no contexto do neoliberalismo. **Revista Mensal**, ano II, nº. 13. Jun. 2002. Disponível em www.espacoacademico.com.br/013. Acesso em março 2003.

BABBIE, Earl. **Métodos de Pesquisas de Survey.** Trad. Guilherme Cezarino. Belo Horizonte: Editora UFMG, 1999.

BAHIA. **Educar para Vencer**. O Ensino Público do Novo Século. *Folder* promocional do programa. Salvador: Governo da Bahia. Secretaria da Educação, 1999.

_____. Escolas: **Projeto de Fortalecimento da Gestão Escolar.** *Folder* promocional do programa. Salvador: Governo da Bahia. Fundação Luis Eduardo Magalhães. Secretaria da Educação, 2000.

_____. **Manual de Gestão Municipal da educação:** Conceitos e Instrumentos. OLIVEIRA, João Batista (coord). Salvador: Secretaria da Educação e Fundação Luis Eduardo Magalhães, 2000.

_____. **Gerenciando a Escola Eficaz: Conceitos e Instrumentos.** OLIVEIRA, João Batista (coord). Salvador: Secretaria da Educação e Fundação Luis Eduardo Magalhães, 2000.

_____.**Projeto de Regularização do Fluxo Escolar 1ª a 4ª série**: transformar a pedagogia da repetência na pedagogia do sucesso. *Folder* promocional do programa. Salvador: Governo da Bahia; Fundação Luis Eduardo Magalhães; Secretaria da Educação, 2000.

_____.**Projeto de Regularização do Fluxo Escolar 5ª a 8ª série:** transformar a pedagogia da repetência na pedagogia do sucesso. *Folder* promocional do programa. Salvador: Governo da Bahia; Fundação Luis Eduardo Magalhães; Secretaria da Educação, 2000.

_____.**Construindo a escola Terra Bahia**. Proposta de Educação Básica. Salvador: Governo da Bahia. Secretária da Educação/SUPEN, 2000.

_____. **Manual do PDE: Orientação para a implantação e implementação**. Salvador: Governo da Bahia; Fundação Luis Eduardo Magalhães; Secretaria da Educação, 2001.

_____. **Curso para gestores**. Salvador: Governo da Bahia; Fundação Luis Eduardo Magalhães; Secretaria da Educação, 2001.

_____. Educar para Vencer. **Projeto de Avaliação Externa**. *Folder* de divulgação. Salvador: SEC/UFBA, 2003.

_____. **Avaliação de Desempenho 2004:** resultados gerais e análises pedagógicas. Relatório da AD 2004. Salvador: Governo da Bahia. Secretaria da Educação, 2005.

BAMBERGER, Michael; RUGH, Jim; MABRY, Linda. **RealWorld Evaluation**: working under budget, time, data, and political constraints. California, Sage Publications, 2006.

BONAMINO, Alicia; BESSA, Nícia; FRANCO, Creso (Org). **Avaliação da Educação Básica**. Rio de Janeiro: Ed. PUC-Rio; São Paulo: Loyola, 2004.

BOORSMA, Peter B. La gerencia pública moderna en la teoria y la práctica. Especial referencia a los Países Bajos. **Revista del CLAD Reforma y Democracia.** Caracas, no. 08, maio 97. Disponível em http://www.clad.org.ve/rev08/0029900.pdf Acesso em 16 dez 2003.

BRASIL. Ministério do Planejamento, Orçamento e Gestão. Secretaria de Gestão. **Gestão pública para um Brasil de todos: um plano para o Governo Lula**. Brasília: MP, SEGES, 2003.

_____. Ministério do Planejamento, Orçamento e Gestão. Secretaria de Gestão. **Gestão Pública Empreendedora.** Brasília: MP/SG, julho 2000

_____. Presidência da República. **Plano Diretor da Reforma do Aparelho do Estado.** Brasília: Câmara da Reforma do Estado, 1995

_____. **PORTARIA Nº 931.** Institui o Sistema de Avaliação da Educação Básica - SAEB, composto por dois processos de avaliação: a Avaliação Nacional da Educação Básica - ANEB, e a Avaliação Nacional do Rendimento Escolar – ANRESC e estabelece suas diretrizes básicas. 21 de março de 2005. Publicada em D.O.U. DE 22/03/2005, P. 17.

_____. **DECRETO Nº. 6.094**. Dispõe sobre a implementação do Plano de Metas Compromisso Todos pela Educação, pela União Federal, em regime de colaboração com Municípios, Distrito Federal e Estados ....pela melhoria da qualidade da educação básica. 24 de Abril de 2007. Publicada em D.O.U. DE 25/04/2007, p. 5

_____. **LEI Nº. 9.394,** de. 20 de dezembro de 1996. Lei de Diretrizes e Bases da Educação Nacional. Dispõe sobre a reforma do sistema educacional brasileiro. Brasília: Diário Oficial da União, Brasília, DF.

_____. **LEI Nº. 9.424,** de. 24 de dezembro de 1996. Dispõe sobre o Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério, na forma prevista no art. 60, § 7º, do Ato das Disposições Constitucionais Transitórias, e dá outras providências Brasília: Diário Oficial da União, Brasília, DF.

_____. **LEI Nº. 10.861**, de 14 de abril de 2004. Institui o Sistema Nacional de Avaliação da Educação Superior e dá outras providências. Brasília: Diário Oficial da União, Brasília, DF.

_____. **LEI Nº. 11.494**, de 20 de junho de 2007. Regulamenta o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação - FUNDEB, de que trata o art. 60 do Ato das Disposições Constitucionais Transitórias; altera a Lei no 10.195, de 14 de fevereiro de 2001; revoga dispositivos das Leis nos 9.424, de 24 de dezembro de 1996, 10.880, de 9 de junho de 2004, e 10.845, de 5 de março de 2004; 20 de junho de 2007. Publicada em D.O.U. DE 21/06/2007, P. 7

_____. **Constituição da República Federativa do Brasil de 1988**, promulgada em 5 de outubro de 1988. Brasília: Diário Oficial da União, Brasília, DF.

BROSE, Markus. O Marco Lógico: instrumento de gestão e comunicação. In: BROSE, M. (Org**). Metodologia Participativa**: uma introdução a 29 instrumentos. Porto Alegre: Tomo Editorial, 2001. p. 279-286.

BUSTELO, Maria. The potential role of *standards* and guidelines in the development of an evaluation culture in Spain. **Evaluation**, v. 12 (4), p.437-453, 2006. Disponível em http://aje.sagepub.com. Acesso em dezembro de 2007.

CALMON, Paulo. Promovendo a utilização da avaliação: uma abordagem baseada na incidência de custos transacionais. In: **Anais do X Congresso Internacional del CLAD sobre la Reforma del Estado y de la Administración Pública**. Santiago, Chile, Out 2005. p 1-14.

CASTANHAR, José Cezar; COSTA, Frederico Lustosa da. Avaliação de programas públicos: desafios conceituais e metodológicos. In: Anais do **VII Congresso Internacional del CLAD sobre la Reforma del Estado y de la Administración Pública,** 08-11 Outubro, 2002, Lisboa, Portugal, CD-ROM. Disponível em http://www.ebape.fgv.br /academico/asp/dsp_professor.asp?cd_pro=33. Acesso em 28 out 2003.

CASTRO, Maria Helena Guimarães de. **A educação para o século XXI**: o desafio da qualidade e da equidade. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais, 1999.

CASTRO, Cláudio de Moura. **Provão**: como entender o que dizem os números. Brasília : INEP, 2001. 23 p. (Série documental. Textos para discussão)

COOK, Thomas D.; GRUDER, Charles L. Metaevaluation research. **Evaluation Review**, v. 2, p.5- 51, Disponível em http://aje.sagepub.com. Acesso em fevereiro 2009.

COOKSY, Leslie j.; CARACELLI, Valerie J. Quality, Context, and Use: Issues in achieving the goals of metaevaluation. **American Journal of Evaluation**, v. 26, p.31-42, 2005. Disponível em http://aje.sagepub.com. Acesso em fevereiro 2008.

CORRALES, Javier. **Aspectos políticos na implementação das reformas educacionais**. Programa de Promoção da Reforma Educativa na América Latina e Caribe – PREAL. nº. 14. 2000. Disponível em www.preal.cl. Acesso em 20 jul 2004.

COSTA, Frederico Lustosa da. Reforma do Estado: restrições e escapismos no funcionamento das "agências autônomas". Programa de estudos e pesquisas em reforma do Estado e governança. **RAP**, Rio de Janeiro 33 (2), 191-199, mar/abr 1999.

_____. Desafios da reforma democrática. In: Anais do VIII Congreso Internacional del CLAD sobre la Reforma del Estado y de la Administración Pública, Panamá, 28-31 Oct. 2003.

_____. Por uma outra reforma do Estado: estratégias alternativas ao paradigma gerencialista: Programa de Estudos e Pesquisas em Reforma do Estado e Governança. **RAP**/ Fundação Getulio Vargas, 2000, p. 267-270. Disponível em http://www.ebape.fgv.br/espanol/academico/asp /dsp_professor.asp?cd_pro=31. Acesso em 28 dez. 2003.

COSTA, Sergio Francisco**. Estatística aplicada à pesquisa em educação**. Brasília: Editora Plano, 2004.

DANTAS, Lys Maria Vinhaes. **Análise da implementação de uma política educacional pioneira na área de avaliação em larga escala na Bahia**. / Lys Maria Vinhaes Dantas. – 2005.Orientador: Prof. Dr. Robert Evan Verhine. Dissertação (mestrado) – Universidade Federal da Bahia. Escola de Administração, 2005. 255 f.

DANTAS, Lys Maria Vinhaes; VERHINE, Robert Evan. Experiência de meta-avaliação na graduação em Pedagogia**. In: 18 EPENN -** Encontro de Pesquisa Educacional do Norte e Nordeste, 2007, Alagoas. **Anais do** 18o Encontro de Pesquisa Educacional do Norte e Nordeste. Maceio - AL : CEDU - UFAL, 2007. v. 01, p. 1-14.

DAVOK, Delsi Fries. **Modelo de meta-avaliação de processos de avaliação da qualidade de cursos de graduação**. Tese de doutorado. Programa de Pós-graduação em Engenharia de Produção. Universidade Federal de Santa Catarina, Orientador Prof. Jair dos Santos Lapa. Florianópolis, 2006,.272 f.

DEMO, Pedro. **Mitologias da Avaliação.** De como ignorar, em vez de enfrentar problemas. 2ª ed. Campinas, SP: Autores Associados, 2002 (Coleção Polêmicas do nosso tempo, 68).

DIAS SOBRINHO, José. **Avaliação**: políticas educacionais e reformas da educação superior. São Paulo: Cortez, 2003.

DYE, Thomas R. **Understanding public policy**. 8th ed. New Jersey-EUA: Prentice Hall, 1995.

ELLIOT, N. G.; FONTANIVE, N. S.; KLEIN, R. A capacitação de professores em avaliação em sala de aula: um esboço de idéias e estratégias. **Ensaio.** Avaliação e Políticas Públicas em Educação. Rio de Janeiro: Fundação Cesgranrio, v. 11, n. 39, abr/jun 2003, p. 141-152.

FARIA, Carlos Aurélio Pimenta de. Idéias, conhecimento e políticas públicas: um inventário sucinto das principais vertentes analíticas recentes. **Revista Brasileira de Ciências Sociais**. Vol. 18, nº. 51, fev. 2003.

FERNANDES, Reynaldo. **Índice de Desenvolvimento da Educação Básica (Ideb**). Brasília: INEP, 2007, Série documental. Textos para discussão. V. 26.

FERRER, Alejandro Tiana. **Tratamiento y usos de la información en evaluación**. Brasília: CESPE, 2002. 2ª Escola Internacional em Avaliação Educacional: análise comparada de sistemas de avaliação. (mimeo). Espana: U.N.E.D, 1997. Disponível em http://www.oei.org.ar/noticias/tratamiento.pdf. Acesso em fevereiro 2009.

FLEURY, Sonia. Reforma administrativa: uma visão crítica. Programa de estudos e pesquisas em reforma do Estado e governança. **RAP**, Rio de Janeiro 31 (4), 299-309, jul/ago 1997.

FORSS, Kim; REBIEN, Claus C.; CARLSSON, Jerker. Process use of evaluation: types of use that precede lessons learned and *feedback*. **Evaluation** 2002 vol. 8, p. 29-45. Disponível em http://aje.sagepub.com. Acesso em fevereiro 2008.

GATTI, Bernardete. O professor e a avaliação em sala de aula. **Estudos de Avaliação Educacional**, São Paulo: Fundação Carlos Chagas, v. 27, jan/jun 2003. p. 97-114

GENTILI, Pablo. Neoliberalismo e educação: manual do usuário. In: SILVA, T.S.; GENTILI, p. (org.) **Escola S.A**: quem perde e quem ganha no Mercado do neoliberalismo. Brasília: CNTE, 1996.

GIMENES, Nelson. Estudo meta avaliativo do processo de auto-avaliação em uma instituição do ensino superior no Brasil. **Estudos em Avaliação Educacional,** v. 18, n. 37, maio/ago. 2007, p. 217-243

GINSBURG, Alan; RHETT, Nancy. Building a better body of evidence: new opportunities to strengthen evaluation utilization. **American Journal of Evaluation** 2003. vol. 24, p. 489-498. Disponível em http://aje.sagepub.com. Acesso em fevereiro 2008.

HARTZ, Zulmira Maria de Araújo. Princípios e padrões em meta avaliação: diretrizes para os programas de saúde. **Ciênc. saúde coletiva,** v.11 n.3 Rio de Janeiro jul./set. 2006 Disponível em http://www.scielo.br/pdf/csc/v11n3/30987.pdf. Acesso em 20.10.07. p. 733-738

HASHIMOTO, Rosa. Certificação Ocupacional em educação: garantia de educação continuada e ascensão salarial. **Anais do VIII Congreso Internacional del CLAD** sobre la reforma del Estado y la Administración Pública, Panamá, out 2003.

HELENE, Otaviano**. O que as avaliações permitem avaliar**. Instituto de Estudos Avançados da Universidade de São Paulo. s/d. Texto disponível em www.iea.usp.br/observatorios/educacao. Acesso em 26.12.07

HENRY, Gary T; MARK, Melvin M. Beyond use: understanding evaluation's influence on attitudes and actions. **American Journal of Evaluation** 2003. vol 24 (3), p. 293-314. Disponível em http://aje.sagepub.com. Acesso em fevereiro 2008.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Relatório de Gestão 2005**. Brasília, fevereiro de 2006. Disponível em http://www.inep.gov.br/download/inep/relatorio_gestao2005.pdf. Acesso em maio de 2009.

_____. **SINAES** – Sistema Nacional de Avaliação da Educação Superior: da concepção à regulamentação / [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira]. – 4. ed., ampl. – Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2007. 224 p.

JOINT COMMITTEE ON *STANDARDS* FOR EDUCATIONAL EVALUATION**. Summary of the student evaluation *standards***. The Evaluation Center. Western Michigan University. 2003. Disponível em http://www.wmich.edu/evalctr/jc/STDStnds-Sum.htm. Acesso em fevereiro de 2008

JOINT COMMITTEE ON *STANDARDS* FOR EDUCATIONAL EVALUATION, the (1994). **The Program Evaluation *Standards***. Thousand Oaks, CA: Sage Publications, Inc. All rights reserved. Approved by the American National *Standards* Institute as an American national *standard*. Approval date: March 15, 1994.

JOINT COMMITTEE ON *STANDARDS* FOR EDUCATIONAL EVALUATION, the (1981). ***Standards* for Evaluations of educational programs, projects, and materials**. NY: McGraw-Hill Book Company, 1981.

LAVILLE, Christian; DIONNE, Jean. **A construção do saber**: manual de metodologia da pesquisa em ciências humanas. Adaptação: Lana Mara Siman. Trad. Heloisa Monteiro e Francisco Settineri. Porto Alegre: Artmed; Belo Horizonte: Editora UFMG, 1999.

LAWRENZ, F.; GULLICKSON, A.; TOAL, S. Dissemination: handmaiden to evaluation use. **American Journal of Evaluation,** 2007. vol. 28, p. 275-289. Disponível em http://aje.sagepub.com. Acesso em fevereiro 2008.

LEEUW, Frans. **Evaluation of development agencies performance**: the role of meta-evaluations. Conference Paper. Fifth Biennial World Bank Conference on Evaluation and Development "Evaluating Development Effectiveness: Challenges and the Way Forward". Washington, D.C. 15-16 July 2003. Disponível em www.worldbank.org/oed/conference2003/papers/leeuw.doc. Acesso em janeiro 2008.

LETICHEVSKY, A. C.; VELLASCO, M. M. B. R, TANSCHEIT, R.; SOUZA, R. C. La Categoría Precisión en la Meta-evaluación: Aspectos Prácticos y Teóricos en un Nuevo Enfoque. **Ensaio:** aval. pol. públ. Educ., Rio de Janeiro, v.13, n.47, p. 255-268, abr./jun. 2005

LETICHEVSKY, Ana Carolina; VELLASCO, Marley M. B.R; TANSCHEIT, Ricardo. Um Sistema Fuzzy de suporte à decisão para meta-avaliação: uma nova abordagem e um estudo de caso desenvolvidos no Brasil. **Ensaio**: aval. pol. públ. Educ., Rio de Janeiro, v.15, n.56, p. 447-462, jul./set. 2007

LEVIN, Jack. **Estatística aplicada a ciências humanas**. 2ª ed. Trad. Sergio Francisco Costa. São Paulo: Editora Harbra. 1987.

LEVITON, Laura C. Evaluation use: advances, challenges and applications. **American Journal of Evaluation** 2003. vol. 24, p. 525-535.

LIBORIO, Helena; COSTA, Jorge Adelino. O Impacto de um programa de avaliação externa no desenvolvimento organizacional de uma escola. **Revista Ensaio: aval. pol. públ. Educ**., Rio de Janeiro, v.12, nº 43, abril/jun. 2004, p.696-710.

LIMA, Licinio C. Modelos organizacionais de escola: perspectivas analíticas, teorias administrativas e o estudo da acção. In: MACHADO, L.M.; FERREIRA, N.S.C (Org.). **Política e Gestão da Educação:** dois olhares. Rio de Janeiro: DP&A, 2002, p. 33-54.

LIPSKY, Michael. **Street –level bureaucracy**: dilemmas of the individual in public services. New York: Russel Sage Foundation, 1980.

LOCATELLI, Isa. Novas Perspectivas de Avaliação. **Revista Ensaio: aval. pol. públ. Educ**., Rio de Janeiro, v.09, nº 33, out/dez. 2001, p.476-488.

LUBISCO, N.M.L; VIEIRA, S.C.; SANTANA, I.V. **Manual de Estilo Acadêmico:** monografias, dissertações e teses. 4ª ed.revista e ampliada. Salvador: EDUFBA, 2008. 145p.

LUCKESI, Cipriano. **Avaliação da Aprendizagem Escolar**. 10ª ed. São Paulo: Cortez, 2000.

MACHADO, Ana Maria Netto. Políticas que impedem o que exigem: dimensões controvertidas na avaliação da pós-graduação. **Universidade e Sociedade**, DF, ano XVI, no. 39, fevereiro de 2007. p. 137-149

MACHADO, Célia Tanajura. **O Banco Mundial e a Educação no Brasil: uma análise comparativa de processos de negociação.** Tese de doutorado. Programa de Pós-graduação em Educação. Faculdade de Educação da Universidade Federal da Bahia, Orientador Prof. Dr. Robert Evan Verhine, 2007.

MARANHÃO. **Diretrizes e Estratégias para a Política Educacional do Estado do Maranhão 1999-2002**. São Luis: Governo do Estado, 2000.

MAY, Henri . Making statistics more meaningful for policy research and program evaluation. **American Journal of Evaluation**, v. 25, p.525-540, 2004. Disponível em http://aje.sagepub.com. Acesso em fevereiro 2008.

McTIGHE, Jay; FERRARA, Steven. **Assessing learning in the classroom**. USA: National Education Association, 1998. Student Assessment Series.

MEDEIROS, Ethel Bauzer. **Medidas psico & lógicas**: introdução à psicometria. Rio de Janeiro: Ediouro, 1999.

MOREIRA, Herivelto. As Perspectivas da Pesquisa Qualitativa para as Políticas Públicas em Educação. **Revista Ensaio**: **aval. pol. públ. Educ**., Rio de Janeiro, v.10, nº 35, abr/jun. 2002.

OLIVEIRA, Cleiton. Gestão da educação: União, Estado/Distrito Federal, município e escola. In: MACHADO, L.M.; FERREIRA, N.S.C (Org.). **Política e Gestão da Educação:** dois olhares. Rio de Janeiro: DP&A, 2002, p. 69-82.

OSBORNE, David. **Reinventando o governo**. Trad. de Sérgio Bath e Ewandro Magalhães Junior. Brasília: MH Comunicações, 1994.

ORTEGÓN, E.; PACHECO, J.F; PRIETO, A. **Metodologia del marco lógico para la planificacion, el seguimiento y la evaluación de proyectos y programas**. CEPAL, Naciones Unidas. Chile, 2005. Serie Manuales no. 42.

PATTON, Michael Quinn. The evaluator's responsibility for utilization**. American Journal of Evaluation.** 1988; vol. 90; Reports on topic areas, p. 5-24 Disponível em http://aje.sagepub.com Acesso em fevereiro 2008

PATTON, Michael Quinn. **Utilization-focused Evaluation.** The New Century Text. 3rd Ed. USA, California: Sage Publications, Inc. 1997.

PATTON, Michael Quinn. The Challenges of making evaluation useful. **Ensaio**: aval. pol. públ. Educ., Rio de Janeiro, v.13, n.46, p. 67-78, jan./mar. 2005

PENNA FIRME, Thereza; LETICHEVSKY, Ana Carolina. O desenvolvimento da capacidade de avaliação no século XXI: enfrentando o desafio através da meta-avaliação. **Ensaio**: aval. pol. públ. Educ., Rio de Janeiro, v.10, n.36, p. 289-300, jul./set. 2002

PEREIRA, Luiz Carlos Bresser. Uma reforma gerencial da administração pública no Brasil. **Revista do Serviço Público**, ano 49, no. 01, jan-mar 1998, p. 5 – 42.

PIMENTA, Carlos César Pimenta. A reforma gerencial do estado brasileiro no contexto das grandes tendências mundiais. **RAP**, Rio de janeiro, 32 (5) 173-99, set/out 1998

QUIVY, Raymond; CAMPENHOUDT, Luc Van. **Manual de Investigação em Ciências Sociais**. 2ª educação. Tradução de João Minhoto Marques, Amália Mendes, Maria Carvalho. Lisboa: Gradiva Publicações, 1998.

RAVELA, P.; ARREGUI, P.; VALVERDE, G.; WOLFE, R.; FERRER, G.; RIZO, F.M; AYLWIN, M.; WOLFF, L. **Las evaluaciones educativas que América Latina necesita**. PREAL, Santiago, 2008. Serie Documentos. No. 40. p. 1-24

RAY, Marilyn. Choosing a Truly External Evaluator**. American Journal of Evaluation.** 2006; vol. 27; p. 372 – 377

REBOLLOSO, E.; FERNÁNDEZ-RAMIREZ, B; CANTON, P; POZO, C. Metaevaluation of a total quality management evaluation system. **Psychology in Spain**, 2002, Vol. 6. No 1, 12-25. Disponível em http://www.psychologyinspain.com/content/full/2002/full.asp?id=6001. Acesso em julho de 2008.

REIS, Nadia Maria Viana. Projeto Fortalecimento da gestão escolar: reflexões sobre os desafios e possibilidades vivenciados na construção progressiva da gestão democrática e da autonomia escolar a partir de uma experiência baiana. In: **Anais do VIII Congreso Internacional del CLAD** sobre la reforma del Estado y la Administración Pública, Panamá, out 2003.

RIBEIRO, Jorge L. Sales. **Características da implementação do Sistema de Avaliação da Educação Superior (SINAES) em instituições de ensino superior**. Orientador: Prof. Dr. Robert Evan Verhine. Tese (doutorado) – Universidade Federal da Bahia. Faculdade de Educação, 2009.

RICKER, Kathryn L. **Setting cut scores**: critical review of Angoff and Modified-Angoff Methods. Centre for Research in Applied Measurement and Evaluation. University of Alberta. 2003. Disponível em http://www.education.ualberta.ca/educ/psych/crame/files/RickerCSSE2003.pdf. Acesso em fevereiro de 2009.

RUA, Maria das Graças. **Análise de políticas públicas:** conceitos básicos. Mimeo [S.l.: s.n.], [ca. 2000].

RUA, Maria das Graças. **A aplicação prática do marco lógico**. s/d Disponível em www.enap.gov.br/downloads/ec43ea4f**Avaliacao**_pratica_marco_logico.pdf. Acesso em 18.03.2008

SANDER, Benno. **Gestão da Educação na América Latina**: construção e reconstrução do conhecimento. Campinas, SP: Autores Associados, 1995.

SANDER, Benno. O estudo da administração da educação na virada do século. In: MACHADO, L.M.; FERREIRA, N.S.C (Org.). **Política e Gestão da Educação:** dois olhares. Rio de Janeiro: DP&A, 2002, p. 55-68.

SANTOS, Reginaldo Souza (Coord**). Esgotamento do Padrão de Financiamento e Crise Fiscal do Estado Brasileiro**. Relatório Preliminar: Projeto de Pesquisa CNPq. Escola de Administração/NPGA. UFBA: Salvador: 2001 (mimeo).

SAUL, Ana Maria. **Avaliação Emancipatória**: desafio à teoria e à prática de avaliação e reformulação de currículo. 6ª ed. São Paulo: Cortez, 2001.

SCHWARTZMAN, Simon. As avaliações de nova geração. In: MELLO E SOUZA (Org). **Dimensões da Avaliação Educacional**. Petrópolis, RJ: Vozes, 2005. p. 15-34

SCRIVEN, Michael. An introduction to meta-evaluation. **Educational Products Report**, v. 2, p. 36-38, 1969. Disponível em http://researcheval.net/metaevaluation.html. Acesso em fevereiro de 2009.

SGUISSARDI, Valdemar. Para avaliar propostas de avaliação do ensino superior. **R. Brás. Est. Pedag.,** Brasília, v. 76, n. 184, p. 549-578, set./dez., 1995

SHULHA, Lyn.M; COUSINS, J. Bradley. Evaluation use: theory, research and practice since 1986. **American Journal of Evaluation**. 1997, vol. 18. p. 195-208.

SRIDHARAN, Sanjeev. Introduction to special section on "What is a useful evaluation?" **American Journal of Evaluation**. 2003, vol. 24. p. 483-487.

SOCIÉTÉ FRANÇAISE DE L´ÉVALUATION. **Charte de l´évaluation des politiques publiques et des programmes publics.** Votada pela Assembléia Geral da SFE em 28 de outubro de 2003. Disponível em www.sfe.asso.fr/docs/*site*/**charte**/sfe_**charte**_2003.pdf. Acesso em fevereiro de 2008.

SOUSA, Sandra M. Zákia L. Possíveis impactos das políticas de avaliação no currículo escolar. **Cadernos de Pesquisa**, n. 119, São Paulo, 2003. p. 175-190. Disponível em www.scielo.br. Acesso em 04.05.2006

SOUZA, Celina. ***Workshop* sobre políticas públicas e avaliação**. Salvador: Fundação Luis Eduardo Magalhães, dez 2002 (mimeo).

SOUZA, Celina. "Estado do Campo" da pesquisa em políticas públicas no Brasil. **Revista Brasileira de Ciências Sociais**. Vol 18, n. 51, fev. 2003.

STAKE, Robert E. The two cultures and the evaluation evolution. **American Journal of Evaluation**. 1982, vol. 3. p. 10-14.

STAKE, Bob. How far dare an evaluator go toward saving the world? **American Journal of Evaluation**. 2004, vol. 25. p. 103-107.

STUFFLEBEAM, Daniel. **Meta-evaluation**. USA: Western Michigan University Evaluation Center. Paper 3. Occasional Paper Series. Dec. 1974. Disponível em www.wmich.edu/evalctr/pubs/ops/ops03.pdf Acesso em julho 2006.

STUFFLEBEAM, D. L. **Program Evaluation Models Metaevaluation *Checklist*** (based on The Program Evaluation *Standards*). USA: Western Michigan University. The Evaluation Center, 1999. Disponível em: http://www.wmich.edu/evalctr/*checklists*/eval_model_metaeval.pdf. Acesso em julho de 2008.

STUFFLEBEAM, D.L.; SHINKFIELD, A.J. **Evaluation theory, models, and applications.** San Francisco, Josey-Baley – Wiley Imprint, 2007.

TEDESCO, J. C. *Educación, Ciudadania y Competitividad em América Latina*. I Encontro Nacional do Fórum Brasil de Educação. Brasília, 18 fev 2003. Coletânea de textos. Disponível em www.portalmec.gov.br/cne/arquivos/pdf/en01_coletaneas.pdf. Acesso em out 2003.

TEIXEIRA, Janssen Edelweiss Nunes Fernandes. **Análise da relação entre a certificação de dirigentes escolares e a implementação do planejamento estratégico nas escolas da rede pública estadual baiana**. Dissertação de Mestrado. Orientador: Prof. Dr. Robert Evan Verhine. Escola de Administração. Universidade Federal da Bahia. - 2006. 199 p.

UNEG. United Nations Evaluation Group. ***Standards* for Evaluation in the UN System**. 2005. Disponível em http://www.uneval.org/papersandpubs/documentdetail.jsp?doc_id=22. Acesso em dezembro de 2008.

UNESCO. **Evaluation Handbook**. Internal Oversight Service. Evaluation Section. IOS/EVS/PI/63. 2007. Disponível em http://unesdoc.unesco.org/images/0015/001557/155748e.pdf. Acesso em junho de 2008.

VIANNA, Heraldo Marelim. Avaliação Educacional: vivência e reflexões. **Revista Estudos em Avaliação Educacional**. São Paulo: Fundação Carlos Chagas, n. 18, jul – dez, 1998, p. 69-109.

_____.(a) Avaliação de sistemas: implementação de políticas públicas. **Revista Estudos em Avaliação Educacional**. São Paulo: Fundação Carlos Chagas, n. 22, jul – dez, 2000, p. 119-133.

_____.(b) **Avaliação Educacional**. Teoria, planejamento, modelos. São Paulo: IBRASA, 2000.

_____. Programas de Avaliação em Larga Escala: algumas considerações. **Revista Estudos em Avaliação Educacional**. São Paulo: Fundação Carlos Chagas, n. 23, jan-jul 2001, p. 93-104.

_____.Avaliações Nacionais em larga escala: análises e propostas. **Revista Estudos em Avaliação Educacional**, no. 27, São Paulo, Fundação Carlos Chagas, jan-jul, 2003, p. 41-76

VERHINE, Robert Evan ; DANTAS, Lys Vinhaes ; SOARES, José Francisco . Do Provão ao ENADE: uma análise comparativa dos exames nacionais utilizados no Ensino Superior. **Ensaio**: aval. pol. públ. Educ., Rio de Janeiro, v. 14, p. 291-309, 2006.

VERHINE, Robert Evan. Avaliação da CAPES: subsídios para a reformulação do modelo. In: MACHADO, D; SILVA Jr, J dos R.; OLIVEIRA, J.F de. (Org.). **Reformas e políticas**: educação superior e pós-graduação no Brasil. Campinas, SP: ALÍNEA, 2008, p. 165-188.

WALKER, ELAINE M. The impact of state policies and actions on local implementation efforts: a study of Whole School Reform in New Jersey. **Educational Policy,** vol. 18, nº 02, May, 2004, p. 338-363.

WEISS, Carol H. Have we learned anything new about the use of evaluation? **American Journal of Evaluation**, vol. 19, n. 01, 1998, p. 21-33 Disponível em http://aje.sagepub.com. Acesso em fevereiro de 2008.

WEISS, Carol H. The interface between evaluation and public policy. **Evaluation,** 1999, vol. 5, p. 468-486. Disponível em http://evi.sagepub.com. Acesso em fevereiro 2008.

WIDMER, Thomas. Evaluating evaluations: does the Swiss Practice live up to the "Program Evaluation *Standards*"? Paper presented at the **I international Evaluation Conference**. Vancouver, Canadá, 2005 p. 67-80. Disponível em www. http://www.seval.ch/documents/unterlagen-*standards*/anwendungen/a17_widmer_1995_evaluating.pdf. Acesso em 26.12.2007

WORTHEN, E; SANDERS, J.R. FITZPATRICK, J.L. **Avaliação de Programas.** Concepções e Práticas. São Paulo: Editora Gente; Edusp; Instituto Ayrton Senna; Instituto Fonte, 2005.

XAVIER, Robina; MEHTA, Amisha; GREGORY, Anne. **Evaluation in use**: the practitioner view of effective evaluation. Queensland University of Technology, Australia. s/d Disponível em: http://praxis.massey.ac.nz/fileadmin/Praxis/Files/Journal_Files/Evaluation_Issue/XAVIER_ET_AL_ARTICLE.pdf. Acesso em 26.12.2007.p. 1-11

YANG, Huilan; SHEN, Jianping. When Is an External Evaluator No Longer External? Reflections on Some Ethical *Issues**. American Journal of Evaluation.*** 2006; vol. 27; p 378-382

YAZBECK, Lola. Sobre avaliação, pesquisas e políticas públicas: considerações de alguns pesquisadores brasileiros. **Ensaio:** aval. pol. públ. Educ., Rio de Janeiro, v.18, n.38, p.9-28, set./dez. 2007

ZABALA, Antoni**. A Prática Educativa.** Como ensinar. Trad. De Ernani F. da. F. Rosa. Porto Alegre: ARTMED, 1998

Apêndice 01

O mapa conceitual que fundamenta a construção do marco teórico para o presente trabalho pode ser visto na ilustração a seguir:

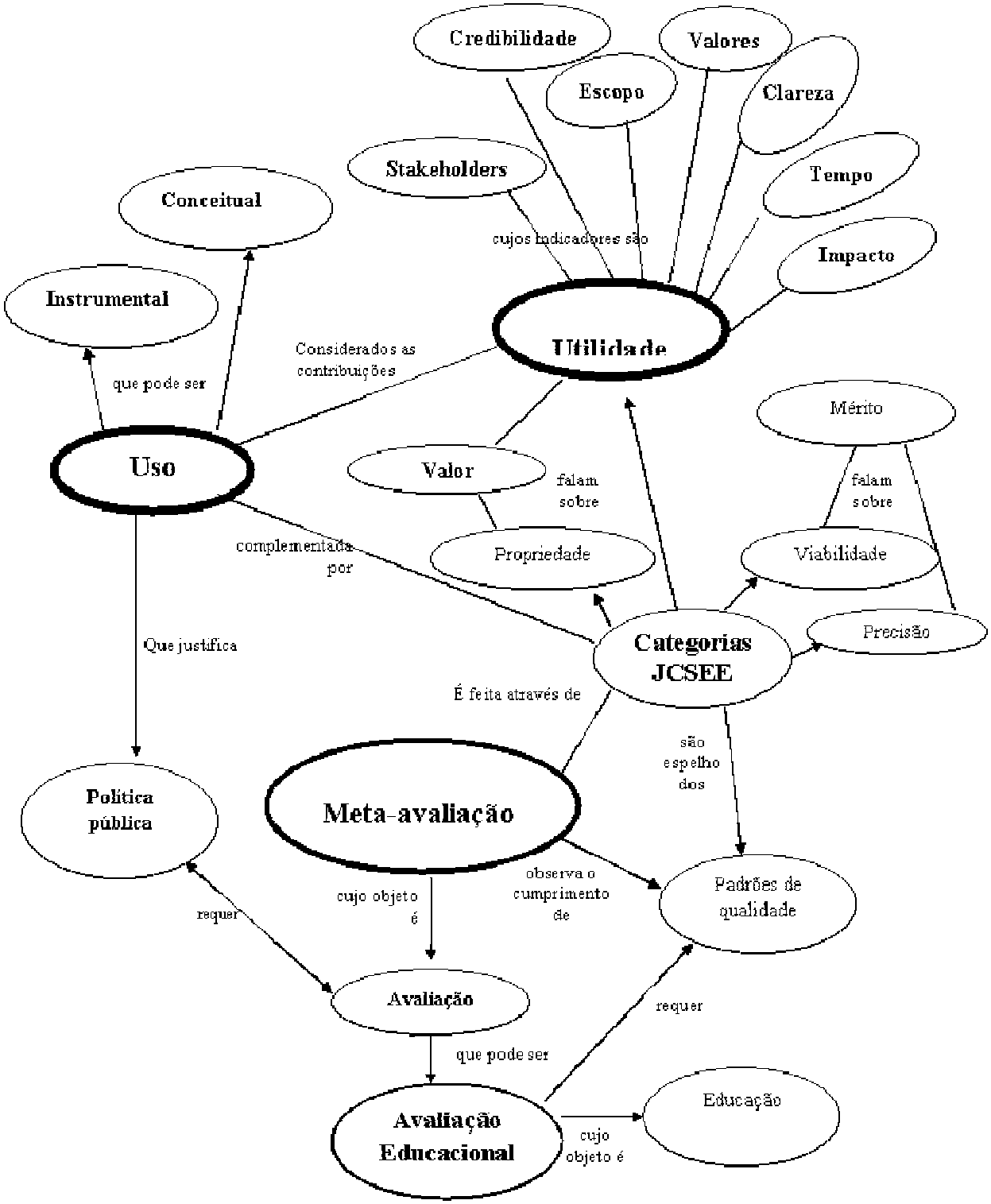


Ilustração 22: Representação do mapa conceitual da tese.

Apêndice 02

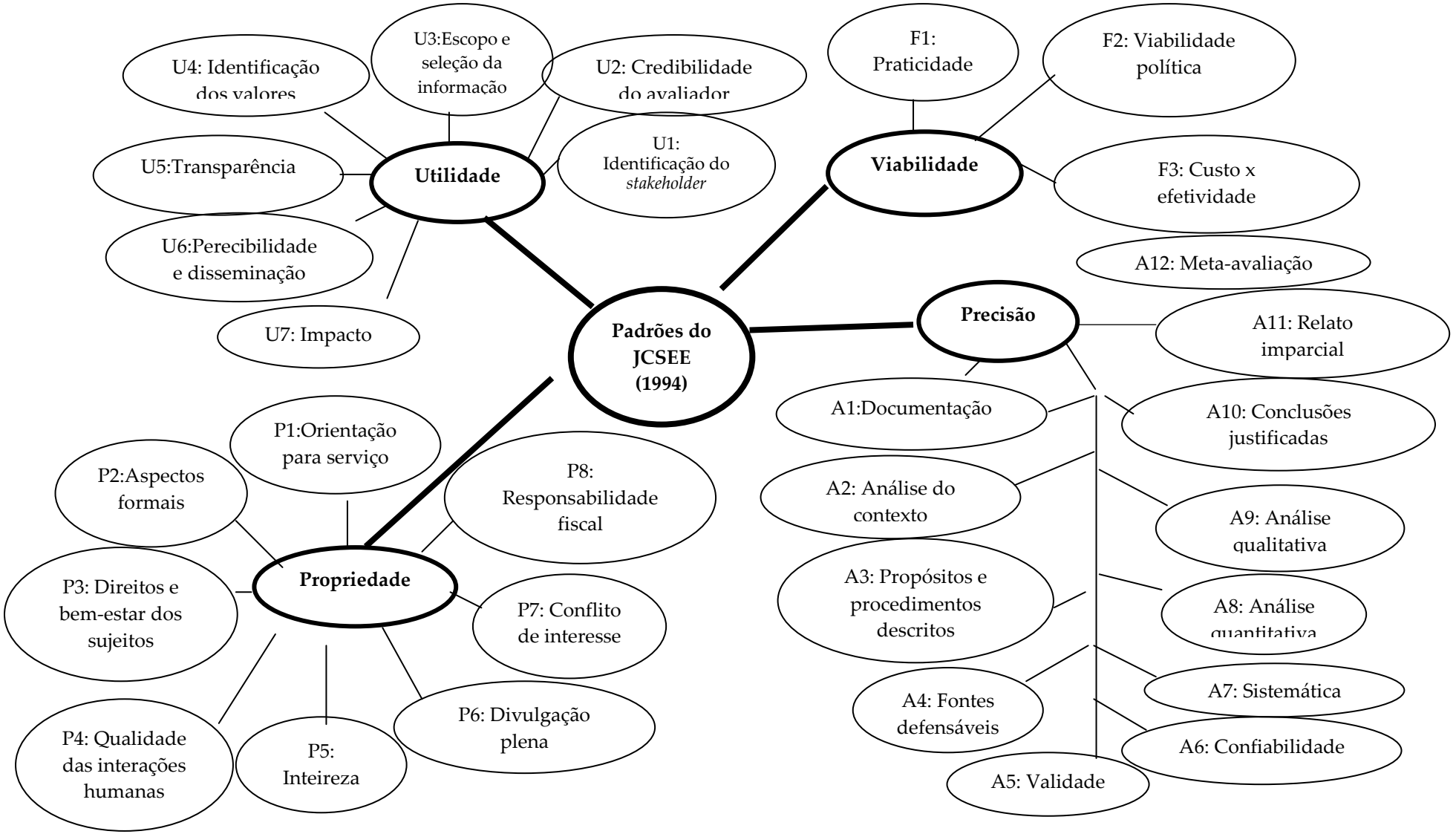


Ilustração 23: Representação esquemática das 4 categorias de padrões do JCSEE, 1994.

Apêndice 03

Relação dos municípios cujos dados foram considerados para as análises de Uso Instrumental

Tabela 64: Freqüência de escolas urbanas e rurais dos municípios cujos dados foram considerados para as análises de Uso Instrumental – dados de 2001 e 2004 – 1ª e 4ª séries do Ensino Fundamental

	Município	Freqüência	Percentual	Percentual
1	ABAIRA	27	,1	,1
2	ABARE	60	,3	,4
3	ACAJUTIBA	20	,1	,5
4	ADUSTINA	38	,2	,7
5	AGUA FRIA	41	,2	,9
6	AIQUARA	16	,1	,9
7	ALAGOINHAS	132	,6	1,5
8	ALCOBACA	59	,3	1,8
9	ALMADINA	22	,1	1,9
10	AMARGOSA	48	,2	2,1
11	AMELIA RODRIGUES	30	,1	2,3
12	AMERICA DOURADA	34	,2	2,4
13	ANAGE	106	,5	2,9
14	ANDARAI	37	,2	3,1
15	ANDORINHA	45	,2	3,3
16	ANGICAL	63	,3	3,6
17	ANGUERA	27	,1	3,7
18	ANTAS	29	,1	3,8
19	ANTONIO CARDOSO	35	,2	4,0
20	ANTONIO GONCALVES	24	,1	4,1
21	APORA	42	,2	4,3
22	APUAREMA	27	,1	4,4
23	ARACAS	31	,1	4,6
24	ARACATU	73	,3	4,9
25	ARACI	145	,7	5,6
26	ARAMARI	28	,1	5,7
27	ARATACA	33	,2	5,8
28	ARATUIPE	39	,2	6,0
29	AURELINO LEAL	32	,1	6,2
30	BAIANOPOLIS	53	,2	6,4
31	BAIXA GRANDE	61	,3	6,7
32	BANZAE	36	,2	6,9
33	BARRA	136	,6	7,5
34	BARRA DA ESTIVA	66	,3	7,8
35	BARRA DO CHOCA	55	,3	8,0
36	BARRA DO MENDES	30	,1	8,2
37	BARRA DO ROCHA	20	,1	8,3
38	BARREIRAS	115	,5	8,8
39	BARRO ALTO	36	,2	9,0
40	BARRO PRETO	20	,1	9,1
41	BARROCAS	37	,2	9,2
42	BELMONTE	50	,2	9,5
43	BELO CAMPO	46	,2	9,7
44	BIRITINGA	46	,2	9,9
45	BOA NOVA	65	,3	10,2
46	BOA VISTA DO TUPIM	85	,4	10,6
47	BOM JESUS DA LAPA	88	,4	11,0
48	BOM JESUS DA SERRA	34	,2	11,1
49	BONINAL	44	,2	11,3

	Município	Freqüência	Percentual	Percentual
50	BONITO	32	,1	11,5
51	BOQUIRA	93	,4	11,9
52	BOTUPORA	47	,2	12,1
53	BREJOES	31	,1	12,3
54	BREJOLANDIA	26	,1	12,4
55	BROTAS DE MACAUBAS	68	,3	12,7
56	BRUMADO	78	,4	13,1
57	BUERAREMA	50	,2	13,3
58	BURITIRAMA	67	,3	13,6
59	CAATIBA	41	,2	13,8
60	CABACEIRAS DO PARAGUACU	24	,1	13,9
61	CACHOEIRA	39	,2	14,1
62	CACULE	58	,3	14,3
63	CAEM	38	,2	14,5
64	CAETANOS	59	,3	14,8
65	CAETITE	142	,7	15,4
66	CAFARNAUM	31	,1	15,6
67	CAIRU	21	,1	15,7
68	CALDEIRAO GRANDE	17	,1	15,8
69	CAMACAN	52	,2	16,0
70	CAMACARI	67	,3	16,3
71	CAMAMU	94	,4	16,7
72	CAMPO ALEGRE DE LOURDES	137	,6	17,4
73	CAMPO FORMOSO	130	,6	18,0
74	CANAPOLIS	34	,2	18,1
75	CANARANA	26	,1	18,2
76	CANAVIEIRAS	64	,3	18,5
77	CANDEAL	32	,1	18,7
78	CANDEIAS	63	,3	19,0
79	CANDIBA	15	,1	19,0
80	CANDIDO SALES	53	,2	19,3
81	CANSANCAO	103	,5	19,8
82	CANUDOS	38	,2	19,9
83	CAPELA DO ALTO ALEGRE	42	,2	20,1
84	CAPIM GROSSO	41	,2	20,3
85	CARAIBAS	83	,4	20,7
86	CARAVELAS	43	,2	20,9
87	CARDEAL DA SILVA	7	,0	20,9
88	CARINHANHA	93	,4	21,3
89	CASA NOVA	260	1,2	22,5
90	CASTRO ALVES	69	,3	22,9
91	CATOLANDIA	14	,1	22,9
92	CATU	61	,3	23,2
93	CATURAMA	39	,2	23,4
94	CENTRAL	48	,2	23,6
95	CHORROCHO	40	,2	23,8
96	CICERO DANTAS	74	,3	24,1
97	CIPO	37	,2	24,3
98	COARACI	33	,2	24,4
99	COCOS	76	,3	24,8
100	CONCEICAO DA FEIRA	41	,2	25,0
101	CONCEICAO DO ALMEIDA	44	,2	25,2
102	CONCEICAO DO COITE	114	,5	25,7
103	CONCEICAO DO JACUIPE	21	,1	25,8
104	CONDE	51	,2	26,0
105	CONDEUBA	62	,3	26,3
106	CONTENDAS DO SINCORA	15	,1	26,4
107	CORACAO DE MARIA	60	,3	26,7
108	CORDEIRO	31	,1	26,8
109	CORIBE	38	,2	27,0

	Município	Freqüência	Percentual	Percentual
110	CORONEL JOAO SA	70	,3	27,3
111	CORRENTINA	65	,3	27,6
112	COTEGIPE	50	,2	27,8
113	CRAVOLANDIA	18	,1	27,9
114	CRISOPOLIS	47	,2	28,1
115	CRISTOPOLIS	42	,2	28,3
116	CRUZ DAS ALMAS	44	,2	28,5
117	CURACA	76	,3	28,9
118	DARIO MEIRA	35	,2	29,0
119	DIAS D'AVILA	25	,1	29,2
120	DOM BASILIO	55	,3	29,4
121	DOM MACEDO COSTA	13	,1	29,5
122	ELISIO MEDRADO	24	,1	29,6
123	ENCRUZILHADA	45	,2	29,8
124	ENTRE RIOS	67	,3	30,1
125	ERICO CARDOSO	46	,2	30,3
126	ESPLANADA	65	,3	30,6
127	EUCLIDES DA CUNHA	159	,7	31,3
128	EUNAPOLIS	53	,2	31,6
129	FATIMA	43	,2	31,8
130	FEIRA DA MATA	22	,1	31,9
131	FEIRA DE SANTANA	280	1,3	33,2
132	FILADELFIA	16	,1	33,2
133	FIRMINO ALVES	10	,0	33,3
134	FLORESTA AZUL	32	,1	33,4
135	FORMOSA DO RIO PRETO	73	,3	33,8
136	GANDU	57	,3	34,0
137	GAVIAO	26	,1	34,2
138	GENTIO DO OURO	52	,2	34,4
139	GLORIA	49	,2	34,6
140	GONGOGI	20	,1	34,7
141	GOVERNAD	20	,1	34,8
142	GOVERNADOR MANGABEIRA	15	,1	34,9
143	GUAJERU	31	,1	35,0
144	GUANAMBI	104	,5	35,5
145	GUARATINGA	113	,5	36,0
146	HELIOPOLIS	48	,2	36,2
147	IACU	64	,3	36,5
148	IBIASSUCE	38	,2	36,7
149	IBICARAI	35	,2	36,9
150	IBICOARA	24	,1	37,0
151	IBICUI	46	,2	37,2
152	IBIPEBA	33	,2	37,3
153	IBIPITANGA	66	,3	37,6
154	IBIQUERA	16	,1	37,7
155	IBIRAPITANGA	80	,4	38,1
156	IBIRAPUA	22	,1	38,2
157	IBIRATAIA	50	,2	38,4
158	IBITIARA	65	,3	38,7
159	IBITITA	31	,1	38,8
160	IBOTIRAMA	52	,2	39,1
161	ICHU	12	,1	39,1
162	IGAPORA	50	,2	39,4
163	IGRAPIUNA	65	,3	39,7
164	IGUAI	86	,4	40,1
165	ILHEUS	129	,6	40,7
166	INHAMBUPE	76	,3	41,0
167	IPECAETA	65	,3	41,3
168	IPIAU	48	,2	41,5
169	IPIRA	139	,6	42,2

	Município	Freqüência	Percentual	Percentual
170	IPUPIARA	26	,1	42,3
171	IRAJUBA	22	,1	42,4
172	IRAMAIA	37	,2	42,6
173	IRAQUARA	51	,2	42,8
174	IRARA	52	,2	43,0
175	IRECE	35	,2	43,2
176	ITABELA	39	,2	43,4
177	ITABERABA	97	,4	43,8
178	ITABUNA	133	,6	44,4
179	ITACARE	69	,3	44,7
180	ITAETE	37	,2	44,9
181	ITAGI	41	,2	45,1
182	ITAGIBA	43	,2	45,3
183	ITAGIMIRIM	13	,1	45,4
184	ITAGUACU DA BAHIA	61	,3	45,6
185	ITAJU DO COLONIA	28	,1	45,8
186	ITAJUIPE	52	,2	46,0
187	ITAMARAJU	107	,5	46,5
188	ITAMARI	30	,1	46,6
189	ITAMBE	55	,3	46,9
190	ITANAGRA	19	,1	47,0
191	ITANHEM	65	,3	47,3
192	ITAPARICA	17	,1	47,4
193	ITAPE	14	,1	47,4
194	ITAPEBI	24	,1	47,5
195	ITAPETINGA	43	,2	47,7
196	ITAPICURU	91	,4	48,1
197	ITAPITANGA	25	,1	48,3
198	ITAQUARA	30	,1	48,4
199	ITARANTIM	39	,2	48,6
200	ITATIM	31	,1	48,7
201	ITIRUCU	23	,1	48,8
202	ITIUBA	136	,6	49,5
203	ITORORO	35	,2	49,6
204	ITUACU	79	,4	50,0
205	ITUBERA	54	,2	50,2
206	IUIU	28	,1	50,4
207	JABORANDI	21	,1	50,4
208	JACARACI	45	,2	50,7
209	JAGUAQUARA	74	,3	51,0
210	JAGUARARI	62	,3	51,3
211	JAGUARIPE	58	,3	51,5
212	JANDAIRA	24	,1	51,7
213	JEQUIE	132	,6	52,3
214	JEREMOABO	116	,5	52,8
215	JIQUIRICA	32	,1	52,9
216	JITAUNA	52	,2	53,2
217	JOAO DOURADO	20	,1	53,3
218	JUAZEIRO	128	,6	53,9
219	JUCURUCU	70	,3	54,2
220	JUSSARA	30	,1	54,3
221	JUSSARI	22	,1	54,4
222	JUSSIAPE	32	,1	54,6
223	LAFAIETE COUTINHO	13	,1	54,6
224	LAGOA REAL	50	,2	54,9
225	LAJE	47	,2	55,1
226	LAJEDAO	11	,1	55,1
227	LAJEDINHO	21	,1	55,2
228	LAJEDO DO TABOCAL	21	,1	55,3
229	LAMARAO	27	,1	55,4

	Município	Freqüência	Percentual	Percentual
230	LAPAO	26	,1	55,6
231	LAURO DE FREITAS	52	,2	55,8
232	LENCOIS	24	,1	55,9
233	LICINIO DE ALMEIDA	52	,2	56,2
234	LIVRAMENTO DO BRUMADO	132	,6	56,8
235	LUIS EDUARDO MAGALHAES	16	,1	56,8
236	MACAJUBA	30	,1	57,0
237	MACARANI	32	,1	57,1
238	MACAUBAS	143	,7	57,8
239	MACURURE	41	,2	58,0
240	MADRE DE DEUS	8	,0	58,0
241	MAETINGA	44	,2	58,2
242	MAIQUINIQUE	19	,1	58,3
243	MAIRI	61	,3	58,6
244	MALHADA	44	,2	58,8
245	MALHADA DE PEDRAS	39	,2	59,0
246	MANOEL VITORINO	63	,3	59,2
247	MANSIDAO	50	,2	59,5
248	MARACAS	39	,2	59,6
249	MARAGOGIPE	87	,4	60,0
250	MARAU	74	,3	60,4
251	MARCIONILIO SOUZA	29	,1	60,5
252	MASCOTE	45	,2	60,7
253	MATA DE SAO JOAO	52	,2	61,0
254	MATINA	49	,2	61,2
255	MEDEIROS NETO	43	,2	61,4
256	MIGUEL CALMON	71	,3	61,7
257	MILAGRES	18	,1	61,8
258	MIRANGABA	50	,2	62,0
259	MIRANTE	26	,1	62,1
260	MONTE SANTO	193	,9	63,0
261	MORPARA	32	,1	63,2
262	MORRO DO CHAPEU	77	,4	63,5
263	MORTUGABA	43	,2	63,7
264	MUCUGE	43	,2	63,9
265	MUCURI	38	,2	64,1
266	MULUNGU DO MORRO	37	,2	64,3
267	MUNDO NOVO	35	,2	64,4
268	MUNIZ FERREIRA	27	,1	64,6
269	MUQUEM DE SAO FRANCISCO	18	,1	64,6
270	MURITIBA	27	,1	64,8
271	MUTUIPE	55	,3	65,0
272	NAZARE	49	,2	65,2
273	NORDESTINA	41	,2	65,4
274	NOVA CANAA	50	,2	65,7
275	NOVA FATIMA	17	,1	65,7
276	NOVA IBIA	33	,2	65,9
277	NOVA ITARANA	12	,1	65,9
278	NOVA REDENCAO	23	,1	66,1
279	NOVA SOURE	58	,3	66,3
280	NOVA VICOSA	31	,1	66,5
281	NOVO HORIZONTE	27	,1	66,6
282	NOVO TRIUNFO	32	,1	66,7
283	OLINDINA	57	,3	67,0
284	OLIVEIRA DOS BREJINHOS	99	,5	67,5
285	OURICANGAS	18	,1	67,5
286	OUROLANDIA	23	,1	67,6
287	PALMAS DE MONTE ALTO	85	,4	68,0
288	PALMEIRAS	26	,1	68,2
289	PARAMIRIM	62	,3	68,4

	Município	Freqüência	Percentual	Percentual
290	PARATINGA	102	,5	68,9
291	PARIPIRANGA	55	,3	69,2
292	PAU BRASIL	30	,1	69,3
293	PAULO AFONSO	124	,6	69,9
294	PEDRAO	24	,1	70,0
295	PEDRO ALEXANDRE	66	,3	70,3
296	PIATA	77	,4	70,6
297	PILAO ARCADO	207	1,0	71,6
298	PINDAI	59	,3	71,9
299	PINDOBACU	30	,1	72,0
300	PINTADAS	35	,2	72,2
301	PIRAI DO NORTE	49	,2	72,4
302	PIRIPA	58	,3	72,6
303	PIRITIBA	42	,2	72,8
304	PLANALTINO	37	,2	73,0
305	PLANALTO	46	,2	73,2
306	POCOES	34	,2	73,4
307	POJUCA	23	,1	73,5
308	PONTO NOVO	50	,2	73,7
309	PORTO SEGURO	96	,4	74,2
310	POTIRAGUA	23	,1	74,3
311	PRADO	59	,3	74,5
312	PRESIDENTE DUTRA	53	,2	74,8
313	PRESIDENTE JANIO QUADROS	68	,3	75,1
314	PRESIDENTE TANCREDO NEVES	65	,3	75,4
315	QUEIMADA	70	,3	75,7
316	QUIJINGUE	84	,4	76,1
317	QUIXABEIRA	29	,1	76,2
318	RAFAEL JAMBEIRO	71	,3	76,6
319	REMANSO	139	,6	77,2
320	RETIROLANDIA	32	,1	77,3
321	RIACHAO DAS NEVES	76	,3	77,7
322	RIACHAO DO JACUIPE	48	,2	77,9
323	RIACHO DE SANTANA	43	,2	78,1
324	RIBEIRA DO AMPARO	58	,3	78,4
325	RIBEIRA DO POMBAL	93	,4	78,8
326	RIBEIRAO DO LARGO	61	,3	79,1
327	RIO DE CONTAS	58	,3	79,3
328	RIO DO ANTONIO	56	,3	79,6
329	RIO DO PIRES	36	,2	79,8
330	RIO REAL	74	,3	80,1
331	RODELAS	13	,1	80,2
332	RUY BARBOSA	63	,3	80,5
333	SALINAS DA MARGARIDA	17	,1	80,5
334	SANTA BARBARA	53	,2	80,8
335	SANTA BRIGIDA	53	,2	81,0
336	SANTA CRUZ CABRALIA	27	,1	81,1
337	SANTA CRUZ DA VITORIA	12	,1	81,2
338	SANTA INES	14	,1	81,3
339	SANTA LUZIA	39	,2	81,4
340	SANTA MARIA DA VITORIA	107	,5	81,9
341	SANTA RITA DE CASSIA	65	,3	82,2
342	SANTA TERESINHA	32	,1	82,4
343	SANTALUZ	83	,4	82,8
344	SANTANA	44	,2	83,0
345	SANTANOPOLIS	28	,1	83,1
346	SANTO AMARO	56	,3	83,4
347	SANTO ANTONIO DE JESUS	66	,3	83,7
348	SANTO ESTEVAO	79	,4	84,0
349	SAO DESIDERIO	91	,4	84,4

	Município	Freqüência	Percentual	Percentual
350	SAO DOMINGOS	16	,1	84,5
351	SAO FELIPE	55	,3	84,8
352	SAO FELIX	20	,1	84,9
353	SAO FELIX DO CORIBE	11	,1	84,9
354	SAO FRANCISCO DO CONDE	32	,1	85,1
355	SAO GABRIEL	48	,2	85,3
356	SAO GONCALO DOS CAMPOS	38	,2	85,4
357	SAO JOSE DA VITORIA	22	,1	85,6
358	SAO JOSE DO JACUIPE	25	,1	85,7
359	SAO MIGUEL DAS MATAS	23	,1	85,8
360	SAO SEBASTIAO DO PASSE	43	,2	86,0
361	SAPEACU	35	,2	86,1
362	SATIRO DIAS	43	,2	86,3
363	SAUBARA	9	,0	86,4
364	SAUDE	39	,2	86,5
365	SEABRA	94	,4	87,0
366	SEBASTIAO LARANJEIRA	32	,1	87,1
367	SENHOR DO BONFIM	78	,4	87,5
368	SENTO SE	80	,4	87,9
369	SERRA DO RAMALHO	45	,2	88,1
370	SERRA DOURADA	65	,3	88,4
371	SERRA PRETA	61	,3	88,6
372	SERRINHA	149	,7	89,3
373	SERROLANDIA	33	,2	89,5
374	SIMOES FILHO	70	,3	89,8
375	SITIO DO MATO	34	,2	90,0
376	SITIO DO QUINTO	31	,1	90,1
377	SOBRADINHO	18	,1	90,2
378	SOUTO SOARES	37	,2	90,3
379	TABOCAS DO BREJO VEL	54	,2	90,6
380	TANHACU	82	,4	91,0
381	TANQUE NOVO	69	,3	91,3
382	TANQUINHO	16	,1	91,4
383	TAPEROA	57	,3	91,6
384	TAPIRAMUTA	31	,1	91,8
385	TEIXEIRA DE FREITAS	62	,3	92,1
386	TEODORO SAMPAIO	14	,1	92,1
387	TEOFILANDIA	61	,3	92,4
388	TEOLANDIA	42	,2	92,6
389	TERRA NOVA	21	,1	92,7
390	TREMEDAL	99	,5	93,1
391	TUCANO	104	,5	93,6
392	UAUA	87	,4	94,0
393	UBAIRA	63	,3	94,3
394	UBAITABA	27	,1	94,4
395	UBATA	38	,2	94,6
396	UIBAI	32	,1	94,8
397	UMBURANA	20	,1	94,8
398	UNA	77	,4	95,2
399	URANDI	61	,3	95,5
400	URUCUCA	76	,3	95,8
401	UTINGA	34	,2	96,0
402	VALENCA	135	,6	96,6
403	VALENTE	42	,2	96,8
404	VARZEA DA ROCA	33	,2	97,0
405	VARZEA DO POCO	16	,1	97,0
406	VARZEA NOVA	15	,1	97,1
407	VARZEDO	29	,1	97,2
408	VERA CRUZ	44	,2	97,4
409	VEREDA	29	,1	97,6

	Município	Freqüência	Percentual	Percentual
410	VITORIA DA CONQUISTA	228	1,0	98,6
411	WAGNER	22	,1	98,7
412	WANDERLEY	54	,2	99,0
413	WENCESLAU GUIMARAES	89	,4	99,4
414	XIQUE-XIQUE	137	,6	100,0
	Total	21.759	100,0	

Fontes: SEC/MEC. Censo Escolar 2001 e Censo Escolar 2004

Os municípios de Jacobina, Nilo Peçanha e Salvador foram excluídos da base de dados para efeitos das análises por terem cumprido um ciclo completo apenas com a rede estadual. Os outros 414 municípios foram mantidos.